Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти перший (бакалаврський)

ПРОГНОЗУВАННЯ РОЗВИТКУ ЦУКРОВОГО ДІАБЕТУ ЗА ДОПОМОГОЮ ШТУЧНОГО ІНТЕЛЕКТУ
(тема)

Виконав:
студент 4 курсу, групи ІТІНФ-20-2

Осика Т. С.
(прізвище, ініціали)

Спеціальності 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник ст. викл. Путятіна О.Є.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Кобилін О.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра Інформатики
(повна назва)

Рівень вищої освіти перший (бакалаврський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Осиці Тарасу Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Прогнозування розвитку цукрового діабету за допомогою штучного інтелекту

затверджена наказом університету від 20 травня 2024 року № 464 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 27 травня 2024 р.

3. Вихідні дані до роботи науково-методична та науково-технічна література, дані інтернет-мережі, мова програмування Python, веб-інтерактивне обчислювальне середовище Jupyter Notebook, інтегроване програмне середовище розробки PyCharm

4. Перелік питань, що потрібно опрацювати в роботі
1. Аналіз предметної області, використання ШІ у медичній практиці
2. Проєктування моделі штучного інтелекту та аналіз вибірки даних
3. Розробка моделі штучного інтелекту, оцінка точності прогнозування та оптимізація моделі

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) Використання штучного інтелекту за-для діагностування діабету, постановка задачі, тестові зображення, графіки візуалізації вибірки даних

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	08.04.2024	
2	Аналіз завдання, підбір літератури	08.04.24-14.04.24	
3	Аналіз літератури з досліджуваної проблеми	14.04.24-18.04.24	
4	Аналіз технічних засобів	20.04.24-26.04.24	
5	Проєктування моделі	26.04.24-12.05.24	
6	Програмна реалізація	12.05.24-23.05.24	
7	Оформлення пояснювальної записки	23.05.24-26.05.24	
8	Перевірка на плагіат	27.05.24	
9	Рецензування	28.05.24	
10	Підготовка презентації та доповіді	29.05.24-02.06.24	
11	Занесення роботи в електронний архів	06.06.24	
12	Попередній захист кваліфікаційної роботи	06.06.24	

Дата видачі завдання 8 квітня 2024 р.

Студент _____
(підпис)

Керівник роботи _____ ст. викл. Путятіна О.Є.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до кваліфікаційної роботи: 58 с., 29 рис., 31 джерело.

ЦУКРОВИЙ ДІАБЕТ, PIMA INDIANS DIABETES DATASET, МАШИННЕ НАВЧАННЯ, K-НАЙБЛИЖЧИХ СУСІДІВ, ПРОГНОЗУВАННЯ ЗАХВОРЮВАНЬ, ОБРОБКА МЕДИЧНИХ ДАНИХ, ОПТИМІЗАЦІЯ ГІПЕРПАРАМЕТРІВ, GRIDSEARCHCV, ROC-КРИВА, АНАЛІЗ ДАНИХ.

Об’єктом роботи є датасет Pima Indians Diabetes, який містить медичні дані жінок племені Піма.

Метою роботи є розробка та валідація прогностичної моделі за допомогою алгоритму *k*-найближчих сусідів (*k*-NN) для визначення ризику розвитку цукрового діабету 2 типу.

Робота почалась з детального аналізу та візуалізації даних для ідентифікації проблемних областей, таких як пропущені значення та асиметрія розподілів. Застосовані методи введення включали заповнення пропущених значень за допомогою медіани або середнього, в залежності від характеру розподілу змінної. Масштабування даних було здійснено для забезпечення однакового впливу всіх ознак на процес навчання моделі.

Результатом роботи стала навчена модель *k*-NN з оптимально налаштованим значенням параметра *k*, що демонструє високу точність і здатність ефективно класифікувати ризик розвитку діабету на основі аналізованих даних.

DIABETES MELLITUS, PIMA INDIANS DIABETES DATASET, MACHINE LEARNING, K-NEAREST NEIGHBORS, DISEASE PREDICTION, MEDICAL DATA PROCESSING, HYPERPARAMETER OPTIMIZATION, GRIDSEARCHCV, ROC CURVE, DATA ANALYSIS.

The object of work of this paper is the Pima Indians Diabetes dataset, which contains medical data of Pima women.

The purpose of the work is to develop and validate a predictive model using the *k*-nearest neighbors (*k*-NN) algorithm to determine the risk of developing type 2 diabetes.

The work began with detailed data analysis and visualization to identify problem areas such as missing values and skewed distributions. The imputation methods used included filling in missing values using the median or mean, depending on the nature of the variable’s distribution. The data was scaled to ensure that all features had an equal impact on the model training process.

The result of the work was a trained *k*-NN model with an optimally tuned value of the *k* parameter, which demonstrates high accuracy and the ability to effectively classify the risk of developing diabetes based on the analyzed data.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

SVM – Support Vector Machine (метод опорних векторів)

ROC – Receiver Operating Characteristic (робоча характеристика приймача)

AUC – Area Under ROC Curve (площа під ROC-кривою)

k-NN – *k*-nearest neighbors (*k*-найближчих сусідів)

ВСТУП

Сучасні технології та штучний інтелект (ШІ) відкривають нові горизонти в області медичних досліджень та лікування хронічних захворювань, таких як цукровий діабет. Цукровий діабет – це серйозне захворювання, що характеризується порушенням метаболізму та високим рівнем глюкози в крові, що може призвести до різноманітних ускладнень, зокрема серцевих, ниркових, а також ураження нервової системи та зору. Тому своєчасне виявлення та прогнозування розвитку цієї хвороби має вирішальне значення для запобігання її ускладнень.

Розвиток методів штучного інтелекту, зокрема машинного навчання та глибокого навчання, відкриває перспективи для створення ефективних інструментів прогнозування розвитку цукрового діабету. Ці методи дозволяють аналізувати великі обсяги медичних даних, включаючи лабораторні аналізи, клінічні показники, інформацію про спосіб життя та генетичні фактори, щоб ідентифікувати осіб з високим ризиком розвитку діабету на ранніх стадіях.

Актуальність роботи полягає у потребі оптимізації медичного обслуговування, підвищенні якості життя пацієнтів і зниженні економічного тягаря, пов’язаного з лікуванням та ускладненнями цукрового діабету. Використання ШІ для прогнозування діабету може значно покращити процеси діагностики, моніторингу та лікування, сприяючи переходу від реактивного лікування до проактивного та персоналізованого підходу в медицині.

Ця робота спрямована на дослідження даних та розробку моделі штучного інтелекту, яка дозволить прогнозувати розвиток цукрового діабету з високою точністю на основі аналізу різноманітних даних пацієнтів. Результати можуть бути використані для розробки нових клінічних рекомендацій та стратегій управління ризиком розвитку цукрового діабету.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Огляд захворювання

1.1.1 Епідеміологія

Цукровий діабет є одним з найбільших глобальних викликів у сфері охорони здоров’я. За даними Всесвітньої Організації Здоров’я, кількість людей, хворих на діабет, стрімко зросла з 108 мільйонів у 1980 році до більш ніж 422 мільйонів у 2014 році [1]. Прогнозується, що ця кількість продовжить зростати, особливо в країнах, що розвиваються. Цукровий діабет значно збільшує ризик розвитку серйозних ускладнень, знижує якість життя і підвищує смертність.

До основних факторів, що сприяють зростанню захворюваності, належать старіння населення, збільшення кількості випадків ожиріння, а також зміни в способі життя, такі як недостатня фізична активність і незбалансоване харчування [2]. Важливою складовою боротьби з епідемією діабету є підвищення обізнаності про фактори ризику і важливість раннього виявлення.

1.1.2 Типи діабету

Цукровий діабет поділяється на три основні типи: діабет 1-го типу, діабет 2-го типу та гестаційний діабет [3]. Діабет 1-го типу виникає, коли імунна система організму знищує клітини підшлункової залози, що виробляють інсулін. Цей тип діабету частіше діагностується у дітей та молодиків, але може розвинутися у будь’якому віці. Діабет 2-го типу, який становить приблизно 90-95% усіх випадків, зазвичай розвивається у дорослих і пов’язаний з інсулінорезистентністю та недостатнім виробництвом інсуліну.

Гестаційний діабет може виникати у жінок під час вагітності та зазвичай проходить після пологів, але збільшує ризик розвитку діабету 2-го типу у майбутньому. Ці типи діабету мають різні причини та методи лікування, що підкреслює необхідність точної діагностики та індивідуалізованого підходу до кожного пацієнта.

1.1.3 Фактори ризику

Цукровий діабет 2-го типу, найпоширеніший серед усіх типів. Він має численні фактори ризику, які можуть впливати на його розвиток. Насамперед, генетика відіграє значну роль. Однак, генетичні фактори взаємодіють з низкою навколишніх умов, зокрема, спосіб життя має вирішальне значення. Фізична неактивність і зайва вага значно збільшують ризик розвитку цього захворювання.

Іншими факторами є вік і етнічна приналежність. Важливим фактором також є дієта: висококалорійна їжа, багата простими вуглеводами та насиченими жирами, може сприяти інсулінорезистентності [4]. Розуміння цих факторів дозволяє ефективніше ідентифікувати осіб із високим ризиком розвитку діабету і здійснювати профілактичні заходи:

– генетика: наявність родичів першого ступеня з діабетом значно збільшує ризик розвитку цієї хвороби. Спадковість відіграє ключову роль у схильності до діабету 1-го і 2-го типів. Генетичні мутації можуть впливати на здатність організму виробляти або використовувати інсулін;

– спосіб життя: фізична недіяльність та нездорове харчування є важливими факторами ризику. Регулярні вправи та здоровий раціон можуть значно знизити ризик розвитку діабету 2-го типу. З іншого боку, високий рівень споживання цукрів та жирів збільшує ймовірність виникнення діабету та подальших ускладнень зв’язаних із ним;

– ожиріння: зайва вага, особливо жирові відкладення в області живота, викликає інсулінорезистентність, що є основним чинником розвитку діабету 2-го типу. Управління вагою може значно знизити ризик розвитку діабету;

– вік: з віком ризик розвитку діабету збільшується, особливо після 45 років. Це пов’язано зі зниженням фізичної активності, зменшенням м’язової маси та змінами у способах метаболізму;

– етнічна приналежність: Деякі етнічні групи, такі як афроамериканці, латиноамериканці, південноазійці, мають більший ризик розвитку діабету. Це може бути пов’язано з генетичними факторами, а також з особливостями культури та доступом до охорони здоров’я.

1.1.4 Ускладнення

Цукровий діабет може призвести до численних серйозних ускладнень, які не лише впливають на якість життя пацієнтів, але й можуть бути життєво небезпечними. Важливість раннього виявлення полягає у запобіганні або мінімізації цих ускладнень:

– серцево-судинні захворювання: діабет значно збільшує ризик розвитку серцево-судинних захворювань, таких як коронарна артеріальна хвороба, інфаркт міокарда, інсульт. Високий рівень глюкози в крові може призводити до пошкодження внутрішньої оболонки кровоносних судин, сприяючи атеросклерозу. Ці умови ускладнюють циркуляцію крові і можуть в кінцевому результаті призвести до серйозних кардіологічних подій, які потребують негайної медичної допомоги;

– діабетична нейропатія: це ушкодження нервових волокон у всьому тілі, яке найчастіше вражає нижні кінцівки. Симптоми включають біль, поколювання, втрату чутливості, що значно знижує якість життя і збільшує

ризик травм через втрату відчуттів. У важких випадках нейропатія може призводити до деформації стопи, що вимагає ортопедичного втручання;

– діабетична ретинопатія: це одна з основних причин втрати зору серед людей у працездатному віці. Високі рівні цукру в крові пошкоджують дрібні кровоносні судини в сітківці, що може призвести до кровотеч, утворення рубців і, в кінцевому підсумку, до відшарування сітківки. Регулярні огляди з офтальмологом та своєчасне лікування лазером або іншими методами можуть допомогти запобігти втраті зору;

– ниркова недостатність: діабет є однією з основних причин хронічної ниркової недостатності. Високі рівні глюкози поступово пошкоджують нирки, зокрема фільтруючі структури, що в кінцевому підсумку може призвести до необхідності діалізу. Раннє виявлення змін у роботі нирок і агресивне управління рівнями цукру та артеріального тиску можуть запобігти або затримати прогресування до кінцевої стадії ниркової хвороби;

– діабетична стопа: це серйозне ускладнення, яке включає ушкодження нервів і поганий кровообіг в ногах, що може призвести до інфекцій, виразок і навіть ампутацій. Пацієнти з діабетом мають регулярно перевіряти свої ноги на наявність травм, тріщин або виразок, а також використовувати спеціальне взуття для запобігання травм. Раннє лікування інфекцій та вчасне хірургічне втручання у випадку виразок можуть запобігти більш серйозним наслідкам.

Регулярний моніторинг рівня глюкози в крові, дотримання дієти, фізична активність та правильний прийом ліків є ключовими аспектами лікування. Крім того, регулярні консультації з лікарями та іншими медичними працівниками дозволяють вчасно виявляти будь-які зміни в стані здоров’я і коригувати терапію. Важливо також приділяти увагу психоемоційному стану, оскільки стрес та депресія можуть негативно впливати на перебіг захворювання.

1.1.5 Лікування та стримування розвитку захворювання

Ефективне стримування розвитку цукрового діабету вимагає інтегрованого підходу, який включає медикаментозне втручання, зміни в способі життя, регулярний самоконтроль та освіту пацієнтів.

Медикаментозне лікування – це основа лікування діабету 1-го типу, де пацієнтам потрібен інсулін, оскільки їх тіло не може його виробляти. Інсулін адмініструється ін’єкціями або за допомогою інсулінових помп, що дозволяє точно контролювати рівень цукру в крові. Для діабету 2-го типу використовуються різні класи ліків, що знижують рівень глюкози в крові, такі як метформін, сульфонілсечовини, інсулін та інші [5]. Важливо індивідуально підбирати лікування для кожного пацієнта, враховуючи його загальний стан здоров’я, вік, спосіб життя та інші медичні стани.

Зміна способу життя включає рекомендації щодо здорового харчування та фізичної активності. Здорове харчування для діабетиків включає обмеження споживання простих вуглеводів, збільшення кількості клітковини у раціоні та збалансований прийом білків і жирів. Фізичні вправи допомагають підтримувати здорову вагу, покращувати інсулінорезистентність та загальний стан здоров’я. Рекомендується не менше 150 хвилин помірних аеробних вправ на тиждень.

Пацієнти з діабетом повинні регулярно моніторити рівень глюкози в крові за допомогою глюкометрів. Це дозволяє пацієнтам відслідковувати вплив харчування, фізичної активності, ліків та стресу на рівень глюкози. Регулярний самоконтроль є критично важливим для уникнення гіпо- та гіперглікемічних станів, які можуть бути небезпечними [5].

Важливо, також, щоб пацієнти були добре освічені щодо всіх аспектів їхнього стану, включаючи те, як управляти діабетом, як впізнати і лікувати гіпо- та гіперглікемію, та як запобігати ускладненням. Ефективні програми освіти діабетиків можуть включати індивідуальне навчання, групові курси, семінари, а також навчальні матеріали та онлайн-ресурси.

Психологічна підтримка є важливим аспектом. Діагноз «цукровий діабет» може спричинити емоційний стрес. Підтримка психічного здоров’я є важливою складовою комплексної допомоги людям з діабетом. Це може включати доступ до психологічної консультації, підтримуючих груп, та інших ресурсів, що допомагають керувати стресом і емоціями.

1.2 Роль штучного інтелекту в медицині

1.2.1 Історія та розвиток

Застосування штучного інтелекту (ШІ) в медицині має глибокі корені, що сягають середини 20-го століття, коли вперше були розроблені концепції та алгоритми, які поклали початок цьому напрямку. Ранні дослідження зосереджувалися на створенні систем, здатних імітувати клінічне мислення лікарів. Однією з перших відомих систем стала програма MYCIN [6], розроблена в 1970-х роках на базі Стенфордського університету. MYCIN використовувала правила для діагностування інфекційних захворювань та рекомендацій щодо лікування антибіотиками, хоча й ніколи не використовувалася в клінічній практиці через обмеження технологій того часу.

З появою потужніших комп’ютерів та розвитком машинного навчання ШІ став інтегруватися в медицину з новою силою. У 1980-х і 1990-х роках на базі штучного інтелекту були створені численні діагностичні та навігаційні системи, які допомагали медичним працівникам приймати рішення, базуючись на великих обсягах даних. Ці системи використовували бази знань, наповнені даними про симптоми, лікування та відгуки пацієнтів [7].

Сучасний етап розвитку ШІ в медицині характеризується використанням складних алгоритмів глибокого навчання, які можуть аналізувати медичні зображення, інтерпретувати медичні записи та навіть прогнозувати потенційні медичні стани на основі генетичної інформації.

Наприклад, системи глибокого навчання, такі як ті, що розроблені Google Health та DeepMind, продемонстрували здатність виявляти діабетичну ретинопатію та інші офтальмологічні захворювання на рівні або навіть краще, ніж кваліфіковані фахівці [7].

Значення ШІ для медицини продовжує зростати завдяки його здатності обробляти та аналізувати великі обсяги даних швидше та точніше, ніж це можливо для людини. Це не тільки підвищує ефективність медичних досліджень та діагностичних процедур, але й сприяє розвитку персоналізованих підходів у лікуванні, відкриваючи нові можливості для профілактики та лікування захворювань на індивідуальному рівні.

1.2.2 Технології та інструменти

Штучний інтелект (ШІ) революціонізує медицину, зокрема через впровадження передових технологій та інструментів, які покращують діагностику, лікування і моніторинг захворювань. Застосування ШІ в медицині охоплює широкий спектр технологій, кожна з яких має свої особливості та сфери застосування.

Машинне навчання є одним з основних компонентів ШІ в медицині. Воно включає алгоритми, які можуть навчатися з досвіду без чіткого програмування для кожної задачі. В медицині машинне навчання використовується для аналізу великих обсягів даних, від результатів лабораторних аналізів до медичних записів. Ці алгоритми здатні ідентифікувати закономірності та відхилення, які можуть бути неочевидними для людського ока.

Глибоке навчання, підкатегорія машинного навчання, включає моделі, що імітують структуру та функціонування людського мозку за допомогою штучних нейронних мереж. Це особливо корисно для обробки та аналізу медичних зображень, таких як рентгенівські, МРТ або УЗД знімки. Глибоке

навчання дозволяє виявляти незначні патологічні зміни на зображеннях, що можуть вказувати на ранні стадії захворювань, як-от рак або серцеві захворювання.

Натуральна обробка мови (NLP) використовується для аналізу медичних записів, що дозволяє перетворити неструктуровану медичну інформацію в структуровану форму, яку можна легко аналізувати і використовувати для підтримки клінічних рішень. NLP може допомогти у виявленні тенденцій у симптомах, лікуванні та результати, а також у стандартизації записів для подальшого аналізу.

Комп'ютерне бачення ще один важливий інструмент ШІ, який застосовується для ідентифікації, класифікації та кількісного аналізу зображень. У медицині це може бути використано для автоматичного виявлення аномалій на медичних знімках, що забезпечує швидшу та точнішу діагностику.

Роботизована хірургія, хоч і не є чисто програмним аспектом ШІ, але використовує алгоритми машинного навчання для управління хірургічними інструментами, що дозволяє проводити операції з більшою точністю та меншими розрізами. Це сприяє швидшому одужанню пацієнтів та зменшенню кількості ускладнень.

Загалом, ці технології та інструменти створюють основу для нових підходів у медицині, збільшуючи ефективність діагностики та лікування, а також забезпечуючи більш персоналізоване медичне обслуговування. Штучний інтелект має потенціал радикально змінити медичну практику, зробивши її більш точною, ефективною та доступною.

1.2.3 Прогнозування та діагностика

Прогнозування та діагностика захворювань з використанням штучного інтелекту (ШІ) розкриває нові можливості для медичної науки. ШІ може

значно підвищити точність діагностичних процедур та здатність передбачати майбутній розвиток здоров’я пацієнтів, засновуючись на аналізі обширних даних [8]:

– точність діагностики: ШІ використовує алгоритми машинного навчання для аналізу медичних зображень, таких як рентгенівські знімки, МРТ та ультразвукові сканування. Системи глибокого навчання, особливо конволюційні нейронні мережі, ефективно розпізнають патологічні зміни на медичних зображеннях, що часто пропускаються людським оком. Алгоритми ШІ можуть виявляти мінімальні відхилення від норми, що дозволяє діагностувати захворювання на ранніх стадіях;

– прогнозування ризику захворювання: моделі штучного інтелекту використовують історію хвороби, генетичну інформацію та життєві звички пацієнта для прогнозування ймовірності розвитку захворювань, таких як цукровий діабет, серцеві захворювання та онкологічні хвороби. Використання біомаркерів, таких як біохімічні індикатори у крові, сприяє точному визначенню ризику. Прогнозувальні моделі дозволяють виявляти осіб з високим ризиком розвитку хвороби до появи симптомів, що може підвищити ефективність профілактичних заході;

– раннє виявлення та втручання: завдяки аналізу медичних даних, ШІ дозволяє виявити мінімальні зміни в здоров’ї, що можуть вказувати на початок захворювання. Раннє втручання на основі даних ШІ може включати рекомендації щодо змін у дієті, фізичних вправах або медикаментозному лікуванні. Прогнозування важких ускладнень у хронічних хворих (наприклад, передбачення гіпоглікемічних станів у пацієнтів з діабетом) може допомогти уникнути небезпечних для життя ситуацій;

– оптимізація медичного обслуговування: ШІ сприяє більш ефективному розподілу медичних ресурсів, наприклад, визначенню потреби у спеціалізованих обстеженнях або втручаннях. Автоматизація рутинних діагностичних процедур знижує навантаження на медичний персонал і

скорочує час очікування для пацієнтів. За допомогою ШІ можна аналізувати ефективність різних лікувальних методик і вибирати найбільш ефективні, на основі великих обсягів клінічних даних.

Це все відкриває нові можливості для медицини, роблячи процеси прогнозування та діагностики не тільки швидшими, але й точнішими, що веде до підвищення загальної якості медичного обслуговування, зменшення витрат і забезпечення кращих результатів для пацієнтів.

1.2.4 Персоналізована медицина

Персоналізована медицина, часто називана медициною, заснованою на даних або точною медициною, розглядається як майбутнє медичної допомоги. Цей підхід використовує детальний аналіз генетичної інформації, біомаркерів та інших особливостей кожної особи для розробки індивідуальних планів лікування та профілактики [9]. Ось детальніше про перспективи та майбутнє персоналізованої медицини:

– індивідуалізовані плани лікування: використання штучного інтелекту в персоналізованій медицині дозволяє лікарям краще розуміти, як різні фактори, такі як генетичний профіль пацієнта, можуть впливати на хворобу та її лікування. Це веде до розробки більш ефективних планів лікування, які можуть знижувати ризик побічних ефектів і покращувати результати лікування;

– геноміка та фармакогенетика: ШІ допомагає аналізувати величезні обсяги геномних даних для ідентифікації мутацій, що впливають на ризик розвитку захворювань. Такий аналіз також може вказувати, як пацієнт відреагує на певні лікарські засоби, дозволяючи тим самим вибирати найбільш ефективні ліки без непотрібних проб та помилок;

– прогнозування реакцій на лікування: застосування машинного навчання для прогнозування реакцій на лікування змінює підходи до ведення пацієнтів. Моделі ШІ можуть передбачити, з якою ймовірністю пацієнти відреагують на лікування певними препаратами, що дозволяє враховувати індивідуальну чутливість до ліків або ймовірність виникнення опірності;

– біоінформатика та інтеграція даних: персоналізована медицина потребує інтеграції різноманітних даних, включаючи генетичні, біомедичні, епідеміологічні та клінічні дані. Біоінформатика відіграє ключову роль у об’єднанні цих даних у цілісні моделі, що дозволяють глибше аналізувати зв’язки між різними типами інформації та підвищувати точність медичних прогнозів;

– етичні та правові виклики: впровадження персоналізованої медицини також стикається з етичними і правовими викликами, зокрема з питаннями конфіденційності та доступу до генетичної інформації. Це вимагає розробки нових нормативних рамок, які б захищали права пацієнтів та водночас сприяли б науковим дослідженням;

– майбутнє та інновації: перспективи персоналізованої медицини виглядають обнадійливо з огляду на швидкий розвиток біотехнологій і штучного інтелекту. У майбутньому можливе створення ще більш точних інструментів для моніторингу та лікування хвороб на індивідуальному рівні, що змінить медичну практику;

– взаємодія з пацієнтами та медичними фахівцями: зростаюче впровадження персоналізованої медицини вимагає нового рівня взаємодії між пацієнтами та лікарями. Медичні працівники потребують нових навичок для інтерпретації складних даних і для спілкування з пацієнтами щодо їхнього здоров’я на основі генетичної інформації та індивідуальних ризиків.

Ці аспекти демонструють, що персоналізована медицина має потенціал радикально змінити медичну практику, роблячи лікування більш

цілеспрямованим та ефективним, водночас створюючи нові виклики і можливості для медичної науки і практики.

1.3 Штучний інтелект та діабет

1.3.1 Виявлення ризиків

Використання штучного інтелекту (ШІ) у виявленні ризиків розвитку цукрового діабету може суттєво змінити підходи до профілактики та ранньої діагностики захворювання. Це стало можливим завдяки застосуванню передових алгоритмів машинного навчання, які аналізують великі набори даних і ідентифікують потенційні ризики до появи симптомів [9].

Алгоритми ШІ використовуються для аналізу генетичних даних, з метою виявлення маркерів, пов’язаних із збільшеним ризиком діабету. Ці генетичні маркери можуть включати мутації або варіанти генів, які були зв’язані з розвитком діабету в наукових дослідженнях.

Скринінг факторів ризику: Алгоритми можуть аналізувати велику кількість факторів ризику, таких як вік, вага, сімейний анамнез, рівні фізичної активності, харчові звички, і попередні медичні умови, такі як гіпертонія або порушення обміну речовин.

Попередні медичні результати: Використання ШІ для аналізу історичних медичних даних пацієнта, включаючи результати аналізів крові на глюкозу і гемоглобін A1c, що може вказувати на попередній ризик розвитку діабету.

Аналіз способу життя: ШІ може обробляти дані про спосіб життя, зібрані через мобільні додатки та інші джерела, щоб визначити, як повсякденні звички можуть впливати на ризик розвитку діабету. Наприклад, низька фізична активність і висококалорійна дієта є відомими факторами ризику.

Моделювання ризиків: Сучасні технології ШІ можуть моделювати різні сценарії на основі наданих даних, щоб допомогти передбачити майбутній

розвиток стану здоров’я на основі поточних трендів і змін у поведінці пацієнтів.

1.3.2 Доступність інструментів ШІ

Доступність інструментів штучного інтелекту (ШІ) в медицині, зокрема в діагностиці та лікуванні цукрового діабету, є критичною проблемою, що впливає на глобальну охорону здоров’я. Незважаючи на значний потенціал ШІ покращувати результати лікування і знижувати витрати, існує ряд викликів, які обмежують широке впровадження цих технологій у клінічну практику, особливо в регіонах з обмеженими ресурсами [10]:

– вартість технологій: одним із головних бар’єрів є висока вартість розробки та імплементації систем ШІ. Розробка ефективних алгоритмів вимагає значних інвестицій, які не всі медичні установи можуть собі дозволити;

– необхідність кваліфікованих фахівців: використання та управління ШІ вимагають наявності кваліфікованих фахівців, таких як інженери з даних, аналітики та медичні інформатики, яких часто бракує, особливо у країнах, що розвиваються;

– інфраструктура даних: ефективне використання ШІ вимагає надійної ІТ-інфраструктури для збору, зберігання та обробки великих обсягів даних. У багатьох регіонах відсутня потрібна ІТ-інфраструктура, що обмежує можливості використання передових аналітичних інструментів;

– юридичні та етичні питання: законодавство, що регулює використання медичних даних та ШІ, варіюється між країнами, що може ускладнити міжнародне співробітництво та обмін технологіями. Крім того, існують питання конфіденційності та зловживання даними;

– освітні бар’єри: медичні працівники потребують додаткової освіти та навчання для ефективного використання ШІ. Необхідність постійного навчання та перенавчання може бути обтяжливою та вимагає часу та ресурсів;

– впровадження у клінічну практику: інтеграція ШІ в клінічну практику вимагає чітких доказів ефективності та безпеки. Клінічні дослідження для валідації інструментів ШІ можуть бути тривалими та дорогими;

– сприйняття технологій: існує скепсис з боку як пацієнтів, так і медичних працівників щодо використання ШІ, що може вплинути на прийняття і впровадження нових технологій. Побоювання щодо втрати особистого контакту між лікарем і пацієнтом також можуть відігравати роль;

– адаптація технологій: необхідність адаптації існуючих систем ШІ до локальних умов та потреб може бути складною. Культурні, мовні та демографічні відмінності вимагають індивідуалізації підходів;

– технічні обмеження: інструменти ШІ можуть мати обмеження, зумовлені недостатньою точністю, проблемами зі збором даних або недостатньою здатністю до загальної адаптації.

Подолання цих викликів вимагає спільних зусиль урядів, освітніх установ, охорони здоров’я та технологічної індустрії для створення доступних, ефективних та безпечних інструментів штучного інтелекту, здатних поліпшити умови лікування та діагностики цукрового діабету на глобальному рівні.

1.3.3 Збір та аналіз даних

Збір та аналіз даних є критичними етапами в процесі використання штучного інтелекту для прогнозування розвитку цукрового діабету. Це процес, який вимагає високої точності та обережності, адже від якості зібраних

даних залежить ефективність подальшого аналізу та точність прогнозів [10, 11].

На початку необхідно визначити, які джерела даних будуть найбільш релевантними для дослідження. Це можуть бути медичні записи, результати лабораторних тестів, дані про спосіб життя пацієнтів, а також інформація з носимих пристроїв, які відстежують здоров'я:

– збір даних: після визначення джерел даних наступним кроком є їх збір. Збір даних повинен бути систематичним і стандартизованим, щоб забезпечити консистентність та відтворюваність результатів;

– перевірка якості даних: критично важливо переконатися, що дані, які використовуються, є точними, повними та актуальними. Неповні або неточні дані можуть призвести до помилок у прогнозуванні та аналізі;

– нормалізація даних: різні джерела можуть надавати дані в різних форматах, тому нормалізація даних є необхідною для їхньої інтеграції. Це включає уніфікацію масштабів вимірювань, перетворення даних у загальний формат та вирішення проблем з несумісністю даних;

– обробка пропущених значень: пропущені дані є звичайною проблемою в медичних даних. Важливо визначити методики для обробки цих пропусків, наприклад, заповнення на основі наявних даних;

– виявлення та обробка викидів: викиди можуть спотворити результати аналізу та прогнозування. Важливо виявити та адекватно обробити викиди, щоб вони не впливали на загальні висновки;

– моделювання: використання алгоритмів машинного навчання для розробки прогнозних моделей. Цей процес включає тренування, тестування та валідацію моделей на зібраних даних;

– оцінка моделі: після розробки моделі необхідно оцінити її продуктивність за допомогою метрик, таких як точність, чутливість, специфічність та площа під кривою ROC;

– ітерація та оптимізація: процес моделювання часто вимагає кількох ітерацій, щоб досконало налаштувати параметри та підібрати найкращі алгоритми для конкретних даних.

1.3.4 Інтеграція в клінічну практику

Інтеграція штучного інтелекту (ШІ) в медичну практику є складним процесом, що вимагає ретельного розгляду нормативних, технічних, освітніх та етичних аспектів. Нормативне регулювання відіграє критичну роль у забезпеченні безпеки та ефективності медичних продуктів на основі ШІ. Це включає сертифікацію нових технологій, що гарантує їхню відповідність стандартам якості та безпеки. Також, необхідно забезпечити, що системи на базі ШІ відповідають всім вимогам до приватності та конфіденційності даних, встановленим у медичній індустрії [12].

Технічна інтеграція вимагає сумісності штучного інтелекту з існуючими ІТ-системами в лікарнях, що може бути викликом через неоднорідність і застарілість деяких систем. Це включає інтеграцію з електронними медичними записами, лабораторними системами, а також з портативними медичними пристроями. Важливо також забезпечити високий рівень захисту даних, щоб запобігти несанкціонованому доступу та іншим ризикам цілісності даних.

Професійне навчання медичних працівників є необхідністю для ефективної роботи з новими технологіями. Лікарям і медсестрам потрібно надати відповідні тренінги для забезпечення правильного розуміння можливостей ШІ, а також навичок інтерпретації та використання результатів, які ці системи надають. Важливо, щоб медичні фахівці були здатні інтегрувати ці дані в клінічний контекст і приймати обґрунтовані рішення на їх основі [12].

Крім технічних та освітніх аспектів, важливо також враховувати етичні моменти, пов’язані з використанням ШІ в клінічній практиці. Це стосується питань конфіденційності, інформованої згоди пацієнтів, а також потенційного

впливу алгоритмічних помилок на здоров’я та благополуччя пацієнтів. Врахування цих аспектів є ключовим для побудови довіри та прийняття ШІ у медицині.

1.4 Постановка задачі

Ця робота зосереджена на використанні методів машинного навчання для прогнозування ризику розвитку цукрового діабету 2 типу на основі датасету Pima Indians Diabetes.

Об’єктом роботи є датасет Pima Indians Diabetes.

Метою роботи є в розробка та тренуванні моделі *k*-найближчих сусідів (*k*-NN), яка здатна точно класифікувати осіб за рівнем ризику розвитку діабету.

Для досягнення цієї мети необхідно виконати:

– комплексне очищення даних, включаючи заповнення пропущених значень та нормалізацію даних, щоб забезпечити надійність результатів;

– здійснити оптимізацію гіперпараметрів моделі, щоб знайти найкраще значення кількості сусідів для *k*-NN, яке максимізує точність прогнозування;

– визначати та проаналізувати основні показники ефективності моделі, таких як матриця помилок, точність, повнота, F1-міра, ROC-крива, та AUC-оцінка.

Це дозволить оцінити здатність моделі правильно класифікувати випадки з високим і низьким ризиком захворіти на діабет.

2 ПРОЄКТУВАННЯ МОДЕЛІ

2.1 Загальний огляд використаних технологій

Для аналізу та прогнозування розвитку цукрового діабету було вибрано набір технологій, які оптимально відповідають задачам дослідження з урахуванням їхньої ефективності, доступності та зручності використання. У цьому розділі розглянемо деталі використання Python і Jupyter Notebook для аналітики даних, бібліотеку Scikit-learn для машинного навчання, а також набір даних Pima Indians Diabetes, який застосовується для тренування та тестування моделей. Кожна з цих технологій відіграє ключову роль у здійсненні дослідження та досягненні цілей проєкту.

2.1.1 Python

Python є однією з найпопулярніших мов програмування на сьогодні, особливо у сферах, де потрібен аналіз даних, штучний інтелект та машинне навчання. Його вибір для проєкту з прогнозування розвитку цукрового діабету зумовлений кількома ключовими факторами. По-перше, Python має простий, легко читаємий синтаксис, що робить його доступним навіть для початківців у програмуванні. Це знижує бар’єр входження і сприяє швидкому розвитку проєктів.

По-друге, Python підтримується великою спільнотою розробників, яка постійно доповнює екосистему новими бібліотеками та інструментами. Це особливо важливо у сфері аналізу даних та машинного навчання, де швидкий розвиток технологій вимагає постійного оновлення знань та інструментів. Бібліотеки, такі як NumPy для обробки масивів даних, pandas для маніпуляції табличними даними, Matplotlib та Seaborn для візуалізації, а також Scikit-learn

для машинного навчання, роблять Python незамінним інструментом в руках дослідника.

Третім важливим аспектом є інтеграційні можливості Python з іншими системами та програмами. Python може взаємодіяти з різними базами даних, вебсервісами та навіть з іншими мовами програмування, що робить його відмінним вибором для інтеграції різноманітних систем в єдине середовище досліджень [13].

Важливим фактором у виборі Python також є його відмінні можливості для роботи з машинним навчанням. Python надає платформи, як TensorFlow і PyTorch, які є лідерами у галузі розробки і навчання глибоких нейронних мереж. Ці платформи дозволяють розробляти складні моделі для прогнозування, класифікації та інших задач машинного навчання з високою точністю та ефективністю [14].

Таким чином, Python є надзвичайно потужним інструментом в руках дослідників, який дозволяє ефективно вирішувати складні задачі в області аналізу даних і машинного навчання.

2.1.2 Jupyter Notebook

Jupyter Notebook є важливим інструментом у сфері аналізу даних та машинного навчання, особливо популярним серед науковців, дата-саєнтистів та аналітиків. Це інтерактивне середовище дозволяє користувачам створювати та обмінюватися документами, що містять живий код, візуалізації, рівняння та текст. Використання Jupyter Notebook у дослідницьких проєктах сприяє гнучкому експериментуванню та спрощенню процесу аналізу даних завдяки можливості виконання коду в інтерактивному режимі.

Однією з ключових переваг Jupyter Notebook є його модульність та підтримка різноманітних мов програмування, включаючи Python, R та Julia. В контексті цього проєкту, Python був обраний як основна мова через свою

популярність та велику кількість бібліотек для обробки даних та машинного навчання. Jupyter підтримує вставку візуалізацій безпосередньо у документ, що дозволяє візуально аналізувати дані та представляти результати у зрозумілій формі. великі обсяги даних, що часто зустрічається в медичних дослідженнях.

Крім того, Jupyter сприяє співпраці та спільній роботі дослідників. Функція спільного доступу через GitHub або інші системи контролю версій дозволяє дослідницьким групам ефективно співпрацювати над складними проєктами, підтримуючи при цьому цілісність та відтворюваність досліджень.

2.1.3 Scikit-learn

Scikit-learn – це одна з найбільш широко використовуваних бібліотек машинного навчання для Python, яка надає різноманітні інструменти для аналізу даних і моделювання. Її популярність серед дослідників та розробників заснована на низці ключових особливостей і переваг, які роблять її ідеальним вибором для багатьох застосувань, включаючи прогнозування розвитку цукрового діабету.

Популярність Scikit-learn зумовлена такими факторами як:

– широка підтримка алгоритмів машинного навчання: Scikit-learn включає широкий спектр алгоритмів, які підходять для різноманітних завдань машинного навчання, включаючи класифікацію, регресію, кластеризацію і зниження розмірності. Такий вибір дозволяє дослідникам підібрати найкращі інструменти для конкретних завдань без потреби в пошуку і інтеграції зовнішніх модулів;

– інтеграція з Python науковим стеком: Scikit-learn легко інтегрується з іншими бібліотеками наукових обчислень Python, такими як NumPy, SciPy і

pandas. Це дозволяє дослідникам легко обробляти дані, проводити розрахунки та інтегрувати машинне навчання в більш широкі дослідницькі процеси;

– документація та спільнота: бібліотека має детальну документацію та широку спільноту користувачів і розробників. Нові користувачі можуть легко знайти ресурси для вирішення проблем, а також приклади коду, що значно спрощує процес навчання та розробки;

– ефективність та масштабованість: Scikit-learn оптимізований для високої продуктивності на Python, з підтримкою многопоточних операцій, що дозволяє ефективно обробляти великі обсяги даних.

Scikit-learn, одна з найпопулярніших бібліотек машинного навчання для Python, вирізняється своїми численними перевагами, хоча має і деякі обмеження. Перш за все, єдність інтерфейсу в Scikit-learn значно спрощує роботу розробників, оскільки всі об’єкти в бібліотеці використовують узгоджений інтерфейс. Це полегшує використання та комбінування різних моделей і методик, дозволяючи дослідникам експериментувати з різними підходами за допомогою стандартних методів навчання, прогнозування та оцінки.

Далі, модульність та розширюваність Scikit-learn роблять цю бібліотеку особливо привабливою для тих, хто шукає гнучкість у своїх проєктах машинного навчання. Бібліотека є набором модулів, які можна налаштовувати незалежно, дозволяючи розробникам включати власні естіматори та адаптувати інтерфейси Scikit-learn під свої унікальні завдання.

Проте, Scikit-learn також має свої недоліки. Наприклад, бібліотека має обмежені можливості для глибокого навчання і не включає інструменти для роботи з деякими передовими моделями глибокого навчання, доступними у таких бібліотеках як TensorFlow чи PyTorch. Це може бути обмеженням для досліджень, що вимагають складних нейронних мереж.

Крім того, Scikit-learn зосереджена на використанні CPU і не підтримує обчислення на GPU, що може бути перешкодою для обробки дуже великих

датасетів або складних моделей, які потребують великої обчислювальної потужності.

2.1.4 Pima Indians Diabetes Dataset

Pima Indians Diabetes Dataset є одним з найбільш часто використовуваних наборів даних у дослідженнях, що стосуються машинного навчання і діабету. Цей набір даних був первісно зібраний Національним Інститутом Діабету, Травлення та Ниркових Захворювань США та містить інформацію про жінок племені Піма, які живуть у районі Фенікс, штат Аризона. Цей датасет включає декілька медичних показників, які були зібрані з метою дослідження впливу спадкових факторів на розвиток цукрового діабету 2-го типу.

Дані складаються з 768 спостережень і 9 змінних, включаючи одну цільову змінну, яка індикатор розвитку діабету у респондентів протягом п’яти років. Основні змінні, які входять до складу цього датасету, включають:

– кількість вагітностей – вказує, скільки разів респондентка була вагітною;

– рівень глюкози в плазмі після 2 годин з моменту проведення перорального тесту на толерантність до глюкози – важливий показник, що використовується для діагностики діабету;

– діастолічний кров’яний тиск (мм рт. ст.) – цей показник відображає тиск у артеріях, коли серце між ударами спочиває;

– товщина шкірної складки трицепса (мм) – вимірюється за допомогою штангенциркуля на верхньому краю руки і використовується для оцінки кількості підшкірного жиру, що може бути індикатором інсулінової резистентності;

– сироватковий інсулін (мкУ/мл) за 2 години після перорального вживання глюкози – цей показник відображає рівень інсуліну, який є гормоном, що регулює рівень глюкози в крові;

– індекс маси тіла (вага в кг/(зріст в м)^2) – цей показник є загальноприйнятим способом оцінки, чи є вага особи нормальною відносно її зросту;

– діабетична родинна функція – показник, що оцінює ймовірність генетичної схильності до діабету на основі сімейної історії;

– вік респондентки на момент збору даних.

Цей набір даних є цінним для дослідників, тому що він дозволяє аналізувати взаємозв’язки між різними біомаркерами та розвитком цукрового діабету. Результати, отримані з цього набору даних, часто використовуються для розробки алгоритмів прогнозування, що можуть визначити ризик розвитку діабету на ранніх стадіях [15].

Важливість цього датасету полягає також у тому, що він дозволяє вченим та клініцистам перевіряти ефективність різних методів лікування та профілактики діабету. Завдяки широкій доступності та високій репрезентативності, Pima Indians Diabetes Dataset є одним із ключових інструментів у медичних дослідженнях, пов’язаних із цукровим діабетом [16].

2.2 Вибір моделі штучного інтелекту

2.2.1 *k*-найближчих сусідів (*k*-NN)

Алгоритм *k*-найближчих сусідів (*k*-NN) – це один із методів навчання, який використовується у машинному навчанні для класифікації та регресії. Він належить до сімейства лінивих алгоритмів навчання та навчання з вчителем. Його основна ідея полягає у використанні подібності між зразками даних: новий зразок класифікується за допомогою голосування більшості або середнього значення кількох найближчих до нього зразків [17].

Ключовим параметром у k-NN є k, кількість найближчих сусідів, які беруть участь у визначенні класу або значення нового зразка.

Для визначення того, які зразки є «найближчими», використовується метрика відстані. Найпоширенішою є Евклідова відстань:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}, \tag{2.1}$$

де p і q - точки у n-вимірному просторі.

Також можуть використовуватися Мангеттенська:

$$d(p,q) = \sum_{i=1}^{n}|p_i - q_i|, \tag{2.2}$$

де p і q є точками у n-вимірному.

У класифікації, якщо більшість із k найближчих сусідів належать до певного класу, новий зразок також буде віднесено до цього класу. У випадку регресії середнє значення або медіана відповідей k найближчих сусідів використовується для прогнозування значення.

Для класифікації:

$$\hat{y} = mode\{v_1, v_2, \dots, v_k\}. \tag{2.3}$$

Для регресії:

$$y = \frac{1}{k}\sum_{j=1}^{k} y_j. \tag{2.4}$$

k-NN не вимагає попереднього навчання моделі у звичному розумінні. Немає потреби в оптимізації параметрів, окрім вибору *k*. Це означає, що всі обчислення відбуваються в момент звернення до моделі з новими даними, що робить його високоадаптивним до змін у вхідних даних, але також може призводити до високих обчислювальних витрат.

Також, слід зазначити, що *k*-NN широко використовується у багатьох сферах, включаючи рекомендаційні системи, медичну діагностику, фінансовий аналіз та багато інших застосувань, де швидкість не є критичною, а точність прогнозів важлива;

Користуючись *k*-NN, дослідники та розробники повинні уважно підходити до вибору *k*, оскільки від цього значення залежить баланс між упередженням та дисперсією у прогнозах моделі. Надто маленьке *k* може призвести до перенавчання (overfitting), де модель чутлива до шуму у даних, тоді як надто велике *k* може призвести до перенавчання (underfitting), де модель стає надто загальною і не здатна точно реагувати на більш тонкі особливості даних [18].

2.2.2 Переваги та недоліки *k*-NN

k-найближчих сусідів (*k*-NN) є популярним методом у машинному навчанні завдяки його простоті та ефективності у багатьох сценаріях. Однією з ключових переваг *k*-NN є його непараметрична природа, що означає відсутність необхідності робити припущення про розподіл даних, що робить його гнучким для використання в різноманітних ситуаціях.

Крім того, алгоритм легко адаптується до рішення як задач класифікації, так і регресії, та є відмінним інструментом для виконання багатокласової класифікації. Однак, серед недоліків *k*-NN варто відзначити високу чутливість до шуму у даних та наявність викидів, що можуть суттєво впливати на результати класифікації, особливо при малому кількості сусідів *k* [19]. Також,

k-NN вимагає значних обчислювальних ресурсів для визначення відстаней між зразками, особливо у великих датасетах, та має високі вимоги до пам’яті, оскільки зберігає всі тренувальні дані в пам’яті. Ці характеристики можуть ускладнювати використання *k*-NN у ситуаціях з обмеженими ресурсами або коли потрібна швидка відповідь в реальному часі.

2.2.3 *k*-NN для Pima Indians Diabetes Dataset

Вибір *k*-NN для аналізу датасету Pima Indians Diabetes був обумовлений декількома факторами. Перш за все, цей алгоритм добре підходить для наборів даних, де взаємовідносини між класами можуть бути складними, але локально згрупованими. Оскільки датасет містить відносно невелику кількість зразків (768 осіб), обчислювальні та пам’ятні витрати залишаються в межах прийнятного. Також, враховуючи медичний контекст, критично важливо використовувати модель, яка не робить сильних припущень про розподіл даних, що є сильною стороною *k*-NN.

Користуючись *k*-NN для аналізу Pima Indians Diabetes Dataset, можна взяти до уваги унікальні аспекти медичних даних, особливо коли йдеться про захворювання, які мають комплексні біомедичні підстави такі як діабет [20]. Медичні дані часто містять складні, не лінійні зв’язки між атрибутами, які можуть бути важливими для визначення ризику або прогресування захворювання. *k*-NN ефективно враховує ці зв’язки без потреби моделювання складних параметричних відносин, що робить його особливо цінним для випадків, де біологічні взаємозв’язки можуть не слідувати відомим або очікуваним розподілам.

Однією з ключових переваг *k*-NN є його непараметричний характер, що означає, що алгоритм не робить жорстких припущень про форму розподілу даних. Це особливо важливо для Pima Indians Diabetes Dataset, де показники, такі як рівень глюкози, кількість вагітностей, індекс маси тіла і т.д., можуть

відображати складні взаємодії та неочікувані закономірності [21]. Ці характеристики роблять *k*-NN ідеальним для ідентифікації підгруп пацієнтів, які можуть мати підвищений ризик захворювання на діабет типу 2.

k-NN розглядає лише найближче локальне середовище кожної точки даних, що робить його особливо чутливим до локальних властивостей даних. У контексті Pima Indians Diabetes Dataset, це може означати здатність краще виявляти тонкі особливості в підмножинах даних, що можуть бути важливими для ранньої діагностики діабету. Наприклад, невеликі відмінності у рівнях глюкози або інсуліну, які можуть не впливати на результати в широкомасштабних або глобальних моделях, можуть бути важливими в місцевому контексті кожного пацієнта [22].

Вибір правильного значення *k* є критично важливим для успіху k-NN. Для Pima Indians Diabetes Dataset, занадто мале *k* може призвести до переосмислення шуму та викидів у даних, тоді як занадто велике *k* може «розмивати» корисні місцеві патерни, роблячи прогнози менш чутливими до фактичних біомедичних особливостей даних. Експерименти з різними значеннями *k*, що включають перехресну валідацію та інші методи перевірки, були використані для знаходження оптимального балансу, що максимізує точність прогнозування і мінімізує помилки.

2.2.4 Альтернативи для Pima Indians Diabetes Dataset

k-NN виявився ефективним у визначенні потенційного ризику діабету в Pima Indians Diabetes Dataset завдяки своїй здатності до врахування комплексних, локальних і нелінійних зв’язків у медичних даних. Ця модель, незважаючи на свої недоліки, такі як чутливість до шуму і великі обчислювальні вимоги, демонструє високу ефективність у сценаріях, де точне

і гнучке розуміння даних пацієнтів може мати значний вплив на результати лікування і прогнозування.

Як альтернативи *k*-NN, можна було б розглянути використання інших алгоритмів машинного навчання, які могли б краще впоратися з певними аспектами даних [23]. Наприклад, логістична регресія могла б бути використана для оцінки ймовірностей ризику захворювання, оскільки вона добре підходить для бінарної класифікації завдань і забезпечує легке інтерпретування результатів.

Також випадкові ліси могли б надати більш структурований підхід до класифікації, з можливістю обробки як числових, так і категорійних даних, забезпечуючи при цьому вищу стійкість до викидів і шуму у даних. Окрім того, методи підтримуючих векторних машин (SVM) також могли б бути взяті до уваги, оскільки вони здатні ефективно розділяти датасети з високою розмірністю та комплексністю.

Кожен з цих методів має свої унікальні переваги та обмеження, і вибір найкращого залежить від специфіки даних, цілей дослідження, а також вимог до точності та інтерпретації моделей у клінічному застосуванні.

3 РОЗРОБКА МОДЕЛІ

3.1 Аналіз даних

Аналіз даних у цьому дослідженні починається з вивчення датасету Pima Indians Diabetes. Важливими кроками є виявлення і виправлення викидів, заповнення відсутніх значень та нормалізація даних, що дозволяє забезпечити кращі умови для навчання моделі.

3.1.1 Огляд даних

Визначення розподілу захворюваності на цукровий діабет у наборі даних Pima Indians Diabetes є ключовим кроком в аналізі та розробці прогнозних моделей [24]. Ілюстрація цього розподілу через стовпчикову діаграму та відсоткову кругову діаграму дозволяє наглядно відобразити взаємовідношення між хворими та здоровими особами у вибірці. Перша (рис. 3.1) чітко показує кількість осіб, у яких діагностовано діабет, порівняно з тими, хто залишається без діагнозу.

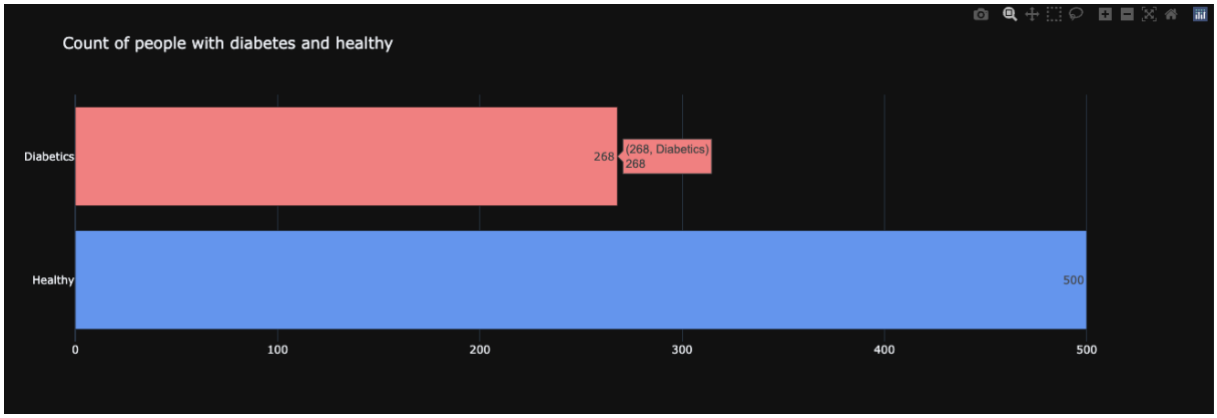

Рисунок 3.1 – Кількість здорових та хворих людей у вибірці

Кругова діаграма (рис. 3.2), в свою чергу, представляє ті ж дані у відсотковому вимірі, ілюструючи частку хворих на діабет від загального числа респондентів. Це дає можливість оцінити відносну пропорційність станів здоров'я у вибірці, що особливо корисно при аналізі впливу захворюваності на популяцію. Відсотки на круговій діаграмі відображаються яскраво та чітко, забезпечуючи легкий доступ до інформації про структуру вибірки.

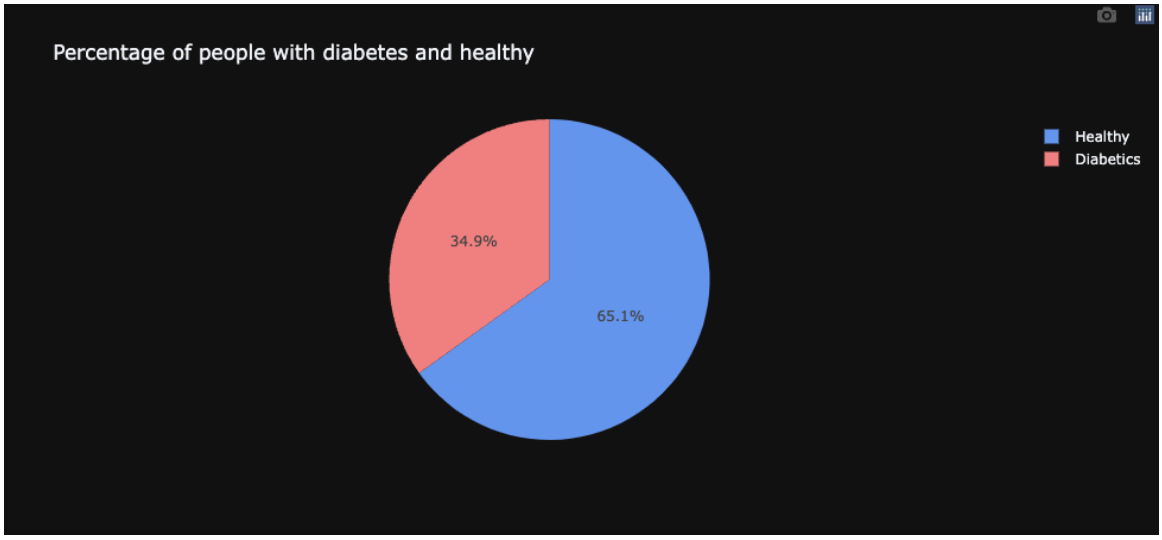

Рисунок 3.2 – Відсоткове відношення здорових та хворих людей

3.1.2 Відсутні дані

Відсутні дані в наборі Pima Indians Diabetes можуть істотно вплинути на якість та достовірність аналітичних висновків, оскільки вони створюють виклики для обробки та аналізу інформації. Проблема відсутніх даних виникає з різних причин, включаючи помилки при зборі інформації, втрату записів або відмову учасників надавати певні дані. Наявність неповних даних вимагає застосування методів вводу, які можуть включати використання середніх значень, медіан або навіть більш складних статистичних технік, таких як багаторазовий ввід або моделювання на основі існуючих патернів у даних. Ці

підходи дозволяють відновити втрачену інформацію та забезпечити більшу точність та об’єктивність результатів дослідження [25].

Для наглядної демонстрації масштабу проблеми з відсутніми даними в дослідженні, планується створення діаграми, яка чітко покаже відсоток відсутності кожної метрики в наборі даних (рис. 3.3). Це допоможе візуально оцінити, які змінні найчастіше мають пропуски, та сприятиме кращому розумінню потенційного впливу цих пропусків на результати дослідження, а також на рішення щодо методів введення та обробки даних.

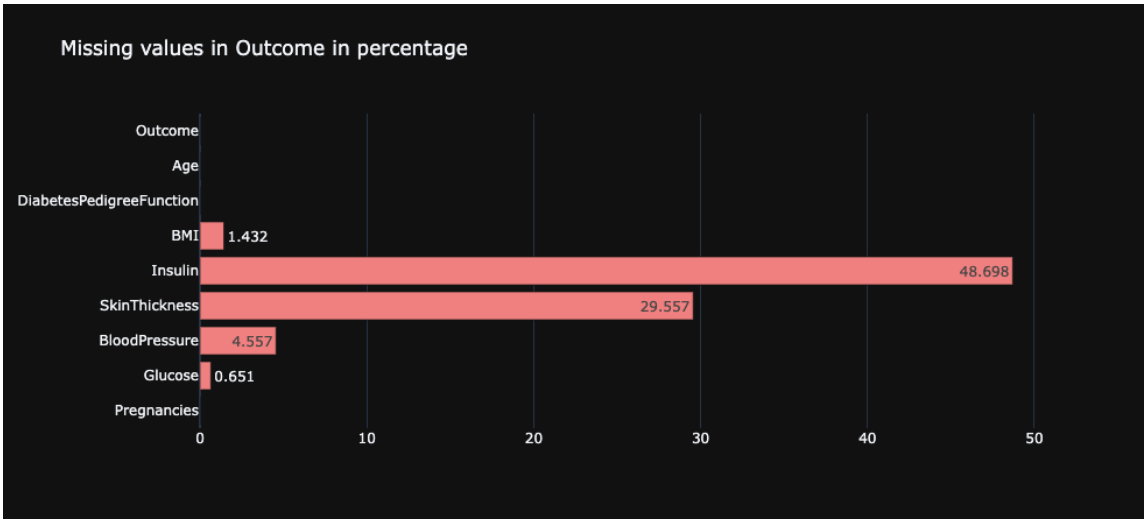

Рисунок 3.3 – Відсоток відсутніх даних

У наборі даних Pima Indians Diabetes деякі ключові метрики мають значні пропуски, що може вплинути на якість аналітичних результатів. Найбільший відсоток пропусків спостерігається у значеннях інсуліну (48,698%), що є критичним показником для оцінки стану діабету, за ним слідує товщина шкірної складки (29,557%), індекс маси тіла (BMI) з пропусками 1,432%, артеріальний тиск (4,557%) та глюкоза (0,651%).

Ці пропуски в даних потребують уважного аналізу та коректного вибору методів, щоб забезпечити точність та надійність наукових висновків.

3.1.3 Матриця кореляцій

Матриця кореляцій – це інструмент, який використовується для визначення та відображення ступеня статистичного зв’язку між різними змінними. Кореляція між двома значеннями може варіюватися від -1 до 1, де 1 означає досконалу пряму кореляцію, -1 – досконалу обернену кореляцію, а 0 означає відсутність лінійного зв’язку. Висока кореляція між двома змінними означає, що коли одна змінна змінюється, друга змінна, ймовірно, зазнає схожих змін в прогнозованому напрямку [26].

У контексті даних Pima Indians Diabetes, може бути виявлено високу кореляцію між такими показниками як рівень глюкози та інсуліну (0,58), що вказує на те, що вищі рівні глюкози в крові часто пов’язані з підвищеним рівнем інсуліну (рис. 3.4). Також висока кореляція між індексом маси тіла та товщиною шкірної складки (0,65), що підкреслює зв’язок між загальною жировою масою тіла та жировими відкладеннями в конкретних областях. Високі кореляційні значення в цих випадках можуть допомогти у виявленні основних факторів, що впливають на розвиток діабету, та сприяти розробці більш ефективних стратегій для його прогнозування та управління [27].

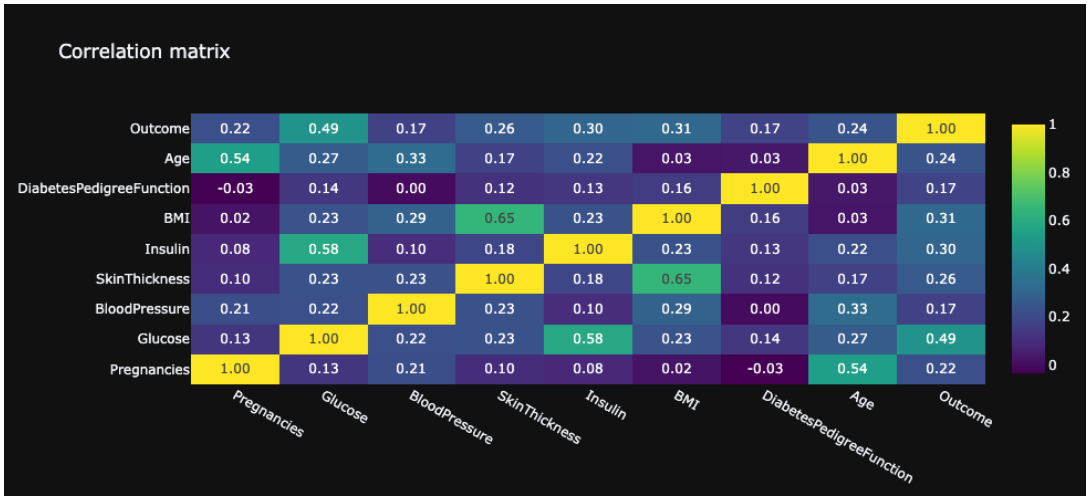

Рисунок 3.4 – Матриця кореляцій для Pima Indians Diabetes

3.2 Підготовка даних

Перед тим, як перейти до побудови моделі, необхідно адекватно підготувати датасет, зокрема заповнити пропущені значення. Цей процес є важливим, оскільки відсутні дані можуть спотворити результати аналізу та знизити точність моделі. Використання ефективних методів допомагає відновити втрачену інформацію та забезпечити більшу консистентність та надійність даних.

3.2.1 Інсулін

У контексті датасету Pima Indians Diabetes, інсулін є критичним показником, який відображає рівень інсуліну в крові після 2 годин пост-глюкозового навантаження. Цей показник має велике значення для оцінки інсулінової резистентності, що часто асоціюється з типом 2 діабету.

Для більш точного розуміння впливу інсуліну на стан здоров'я осіб з діабетом та без, було створено графік, який візуалізує розподілення рівнів інсуліну (рис. 3.5).

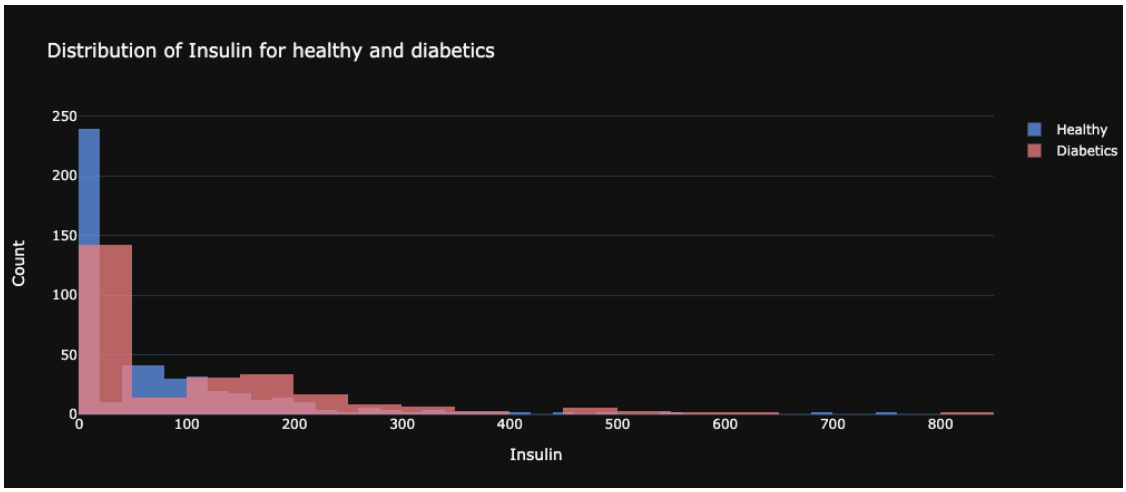

Рисунок 3.5 – Розподілення здорових та хворих людей в залежності від рівня інсуліну

Крім того, було визначено медіанні значення для ненульових вимірів інсуліну, які становили 102,5 для здорової особи та 169,5 для особи з діабетом (рис. 3.6). Це підтверджує, що рівень інсуліну у хворих на діабет вищий, що може свідчити про наявність інсулінової резистентності та необхідність подальшого медичного втручання або моніторингу.

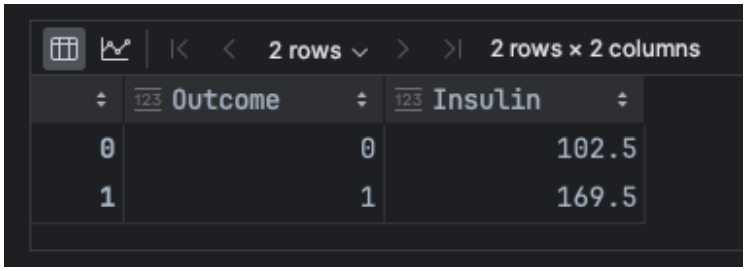

Рисунок 3.6 – Медіанна значення інсуліну для хворих (0) та здорових (1)

Аналіз гістограми рівнів інсуліну в датасеті дозволив виявити, що розподіл є право-асиметричним. Це означає, що більшість значень зосереджені вище медіани, і є довгий хвіст менших значень, який тягнеться вліво. Ліво-асиметричний розподіл часто вказує на те, що менші значення інсуліну є більш поширеними, але також існують рідкісні випадки з значно вищими рівнями (рис. 3.7).

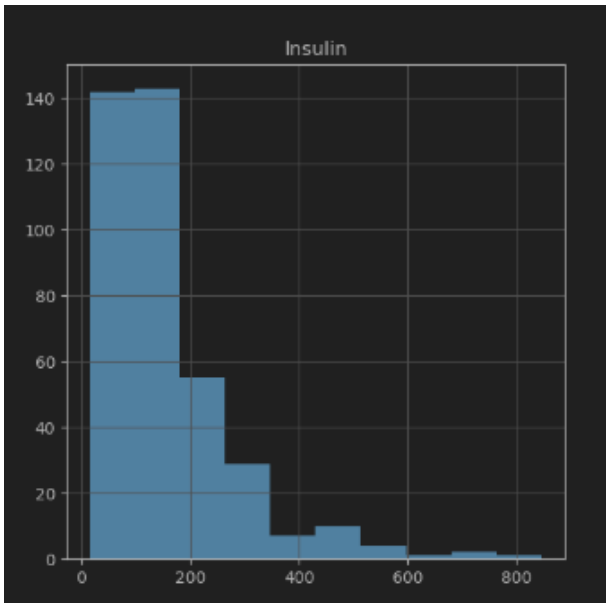

Рисунок 3.7 – Гістограма рівнів інсуліну

У випадках, коли розподіл даних має асиметрію, заповнення відсутніх даних за допомогою середнього значення може бути не найкращим рішенням, оскільки воно може бути зміщеним через наявність викидів. Замість цього, використання медіани (рис. 3.8) для заповнення відсутніх значень є більш доцільним, оскільки медіана стійка до впливу викидів та краще відображає «центральне» значення у випадку асиметричних розподілів.

```
filled_data['Insulin'].fillna(filled_data['Insulin'].median(), inplace = True)
Executed at 2024.05.06 20:14:52 in 3ms
```

Рисунок 3.8 – Заміна пропущених значень за допомогою медіани

3.2.2 Глюкоза

Глюкоза відіграє ключову роль, оскільки це основний індикатор, який використовується для діагностики цукрового діабету. Рівень глюкози в крові після нічного голодування є важливим показником. Високі рівні глюкози можуть вказувати на порушення цього процесу, що є типовим для діабету (рис. 3.9).

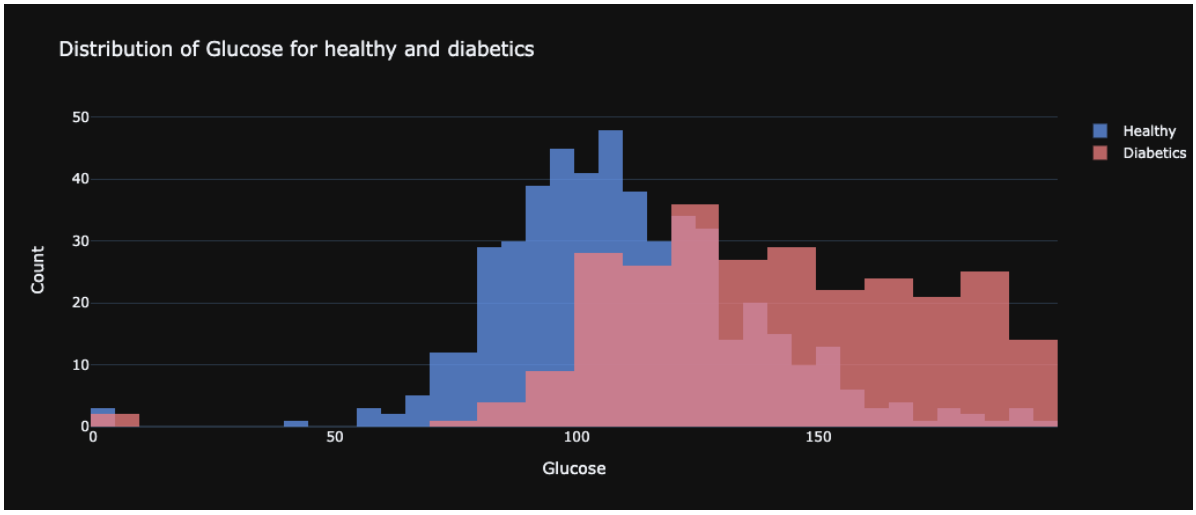

Рисунок 3.9 – Розподілення в залежності від рівня глюкози

Аналіз рівнів глюкози, подібно до аналізу інсуліну, включає вивчення її розподілу серед учасників дослідження. Медіанні значення рівня глюкози склали 107 для здорових осіб та 140 для осіб з діабетом (рис. 3.10). Ці дані підтверджують, що особи з діабетом мають вищий базовий рівень глюкози, що свідчить про метаболічні порушення пов’язані з хворобою [28].

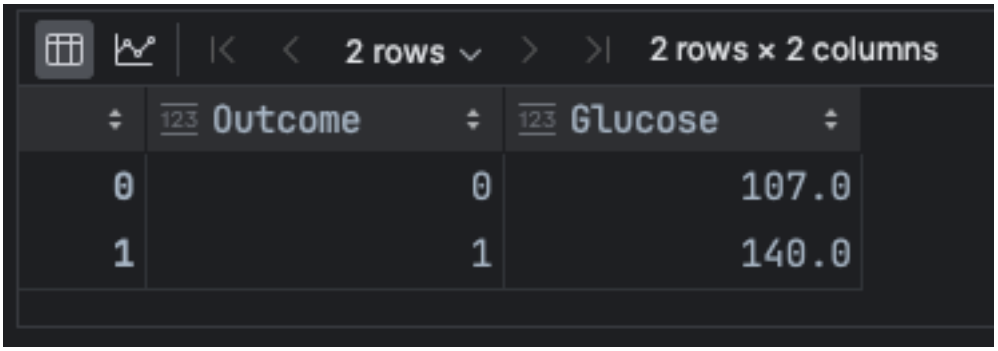

Рисунок 3.10 – Медіанна значення глюкози для хворих (0) та здорових (1)

Важливим аспектом аналізу є також визначення форми розподілу глюкози. В даному випадку, маємо, що розподіл глюкози не має вираженої асиметрії (рис. 3.11), що вказує на більш рівномірне розподілення значень.

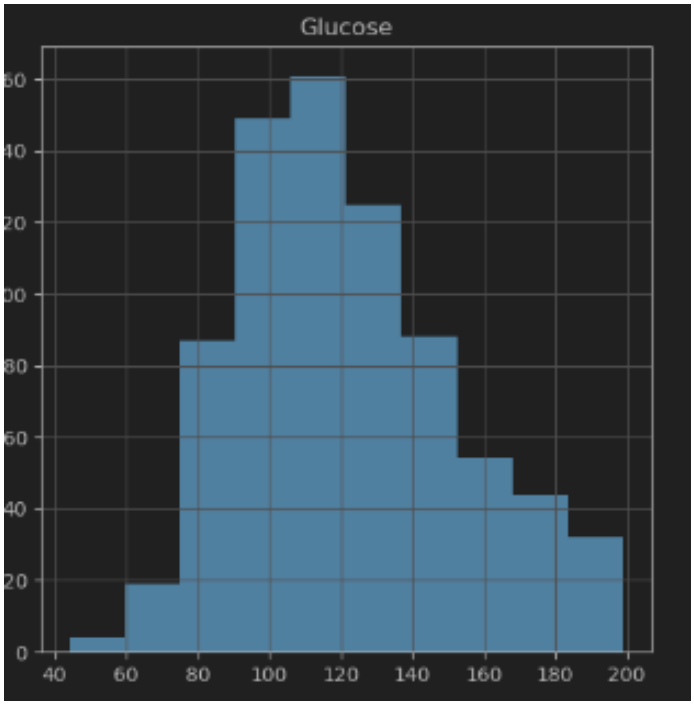

Рисунок 3.11 – Гістограма рівнів глюкози

У таких умовах використання середнього значення для заповнення відсутніх даних є доцільним (рис. 3.12), оскільки середнє надає точну оцінку центральної тенденції розподілу без зміщення, яке могло б виникнути внаслідок асиметрії [29]. Таким чином, для введення відсутніх значень глюкози використовується середнє, що забезпечує збереження внутрішньої структури даних і сприяє точності подальших аналізів і моделювання.

```
'Glucose': filled_data['Glucose'].mean(),
```

Рисунок 3.12 – Заміна значень глюкози на середнє

3.2.3 Товщина шкірної складки трицепса

Товщина шкірної складки трицепса є одним з антропометричних показників, який використовується для оцінки рівня підшкірного жиру в організмі. Цей показник є важливим, оскільки він може вказувати на збільшення ризику розвитку інсулінової резистентності та цукрового діабету 2 типу, особливо якщо він перевищує норму (рис. 3.13).

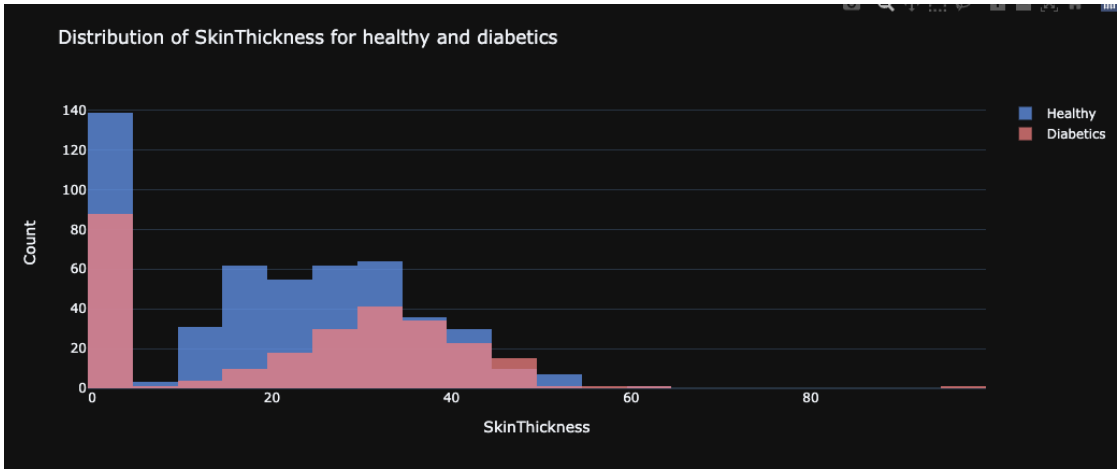

Рисунок 3.13 – Розподілення значень товщини шкірної складки трицепса

В наборі даних Pima Indians Diabetes, медіанні значення товщини шкірної складки трицепса становлять 27 мм для здорових осіб та 32 мм для осіб з діабетом, що відображає загальну тенденцію до вищих значень у групі хворих на діабет (рис. 3.14).

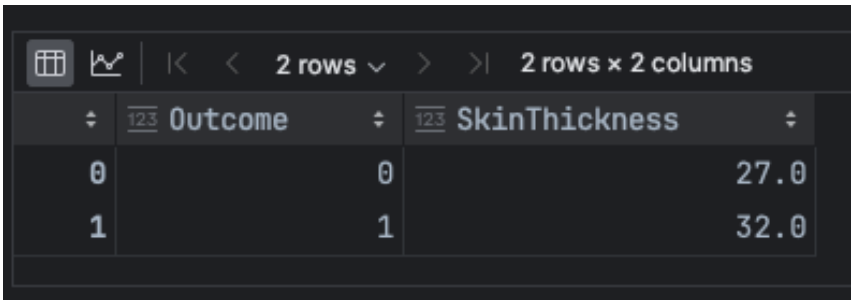

Рисунок 3.14 – Медіана значень товщини шкірної складки трицепса

Аналіз розподілу цього показника показав, що дані мають правосторонню асиметрію (right-skewed), тобто більшість значень зосереджені на нижньому кінці шкали, але існує значна кількість високих значень, що тягнуть середнє у бік більших чисел (рис. 3.15). У зв’язку з цим для заповнення відсутніх значень було обрано використання медіану.

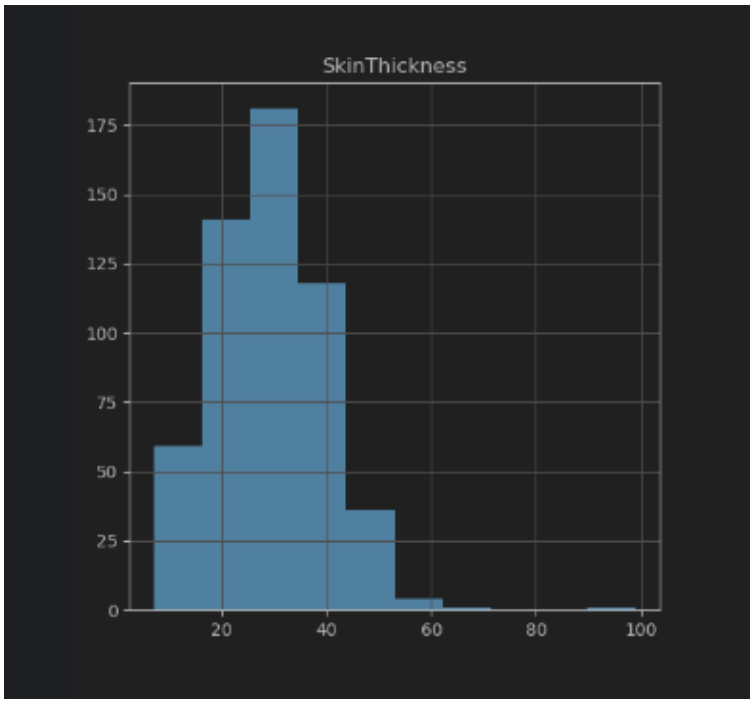

Рисунок 3.15 – Гістограма товщини шкірної складки трицепса

3.2.4 Тиск

Артеріальний тиск у контексті датасету Pima Indians Diabetes відіграє значущу роль, оскільки високий кров'яний тиск є одним з факторів ризику розвитку діабету типу 2 і пов'язаних з ним ускладнень (рис. 3.16).

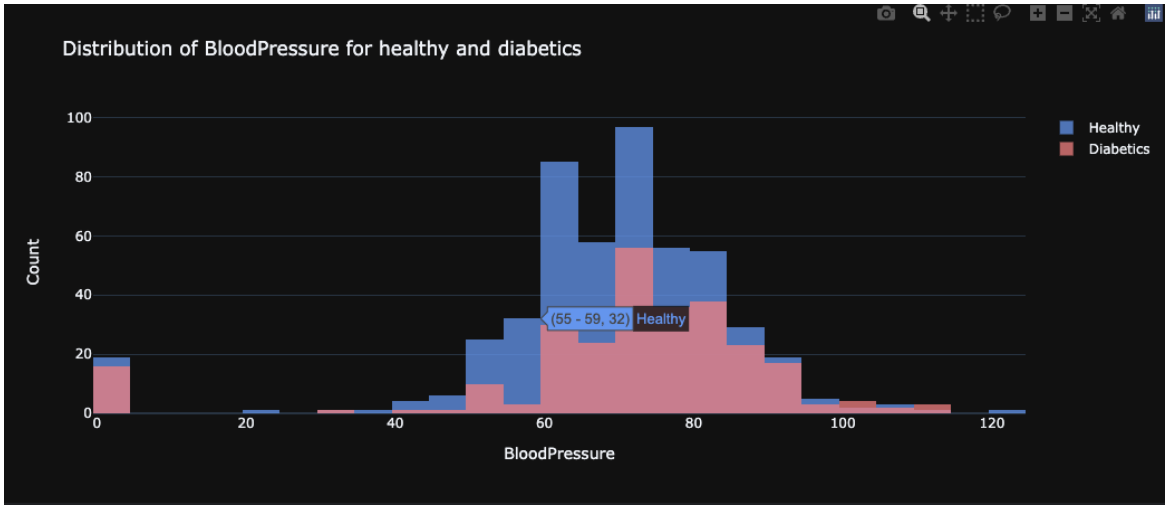

Рисунок 3.16 – Розподілення артеріального тиску

У цьому дослідженні аналіз рівнів артеріального тиску показав, що середні значення становлять 70 мм рт. ст. для здорових осіб та 74,5 мм рт. ст. для осіб з діабетом, що вказує на злегка вищий тиск у групі з діабетом (рис. 3.17).

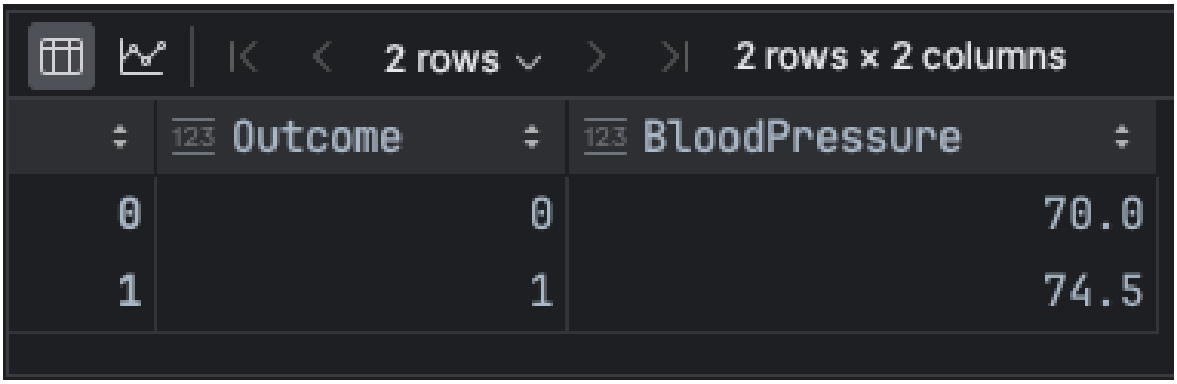

Рисунок 3.17 – Медіана значень артеріального тиску

Для визначення розподілу артеріального тиску в аналізованих даних було встановлено, що вони не мають явної асиметрії. Це означає, що дані розподілені більш-менш рівномірно навколо середнього значення, без виражених хвостів високих або низьких значень (рис. 3.18).

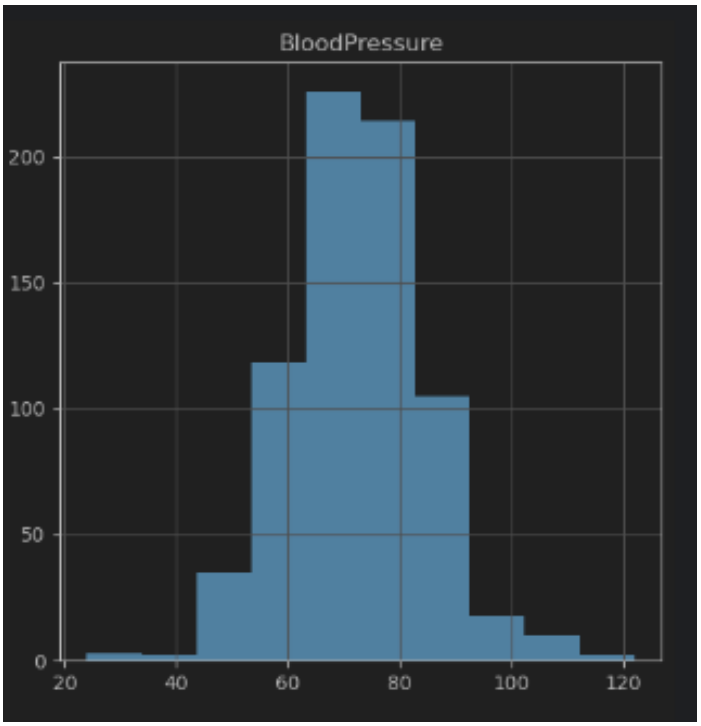

Рисунок 3.18 – Гістограма розподілення значень артеріального тиску

Тому, для заповнення відсутніх даних використовується середнє значення, оскільки воно точно відображає центральну тенденцію у випадку нормального розподілу, забезпечуючи об’єктивність і точність відновлення відсутніх значень.

3.2.5 Індекс маси тіла

Індекс маси тіла (ІМТ) є важливим показником у контексті дослідження, оскільки він допомагає оцінити фізичний стан особи та ризик розвитку діабету (рис. 3.19).

ІМТ розраховується як відношення маси тіла (у кілограмах) до квадрата зросту (у метрах). За значеннями ІМТ можна визначити, чи знаходиться особа у нормальному діапазоні маси тіла, чи має недостатню вагу, надлишкову вагу або ожиріння.

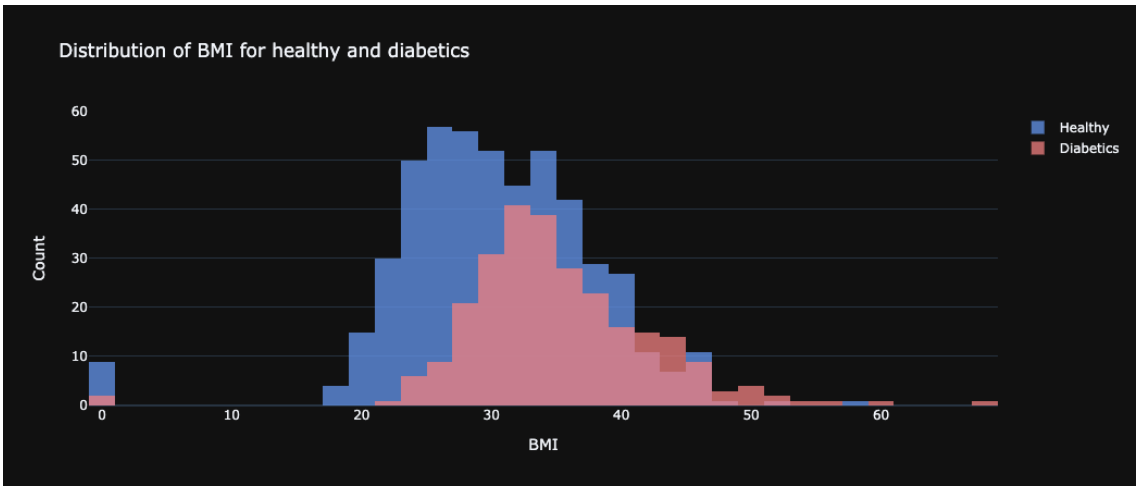

Рисунок 3.19 – Розподілення значень індексу маси тіла

Для оцінки ризику розвитку діабету особливо важливо звертати увагу на надмірну вагу та ожиріння, оскільки ці стани значно підвищують ризик виникнення інсулінорезистентності та, як наслідок, розвитку цукрового діабету 2 типу.

У даному дослідженні, середні значення ІМТ становили 30,1 для здорових осіб та 34,3 для осіб з діабетом, що підтверджує зв’язок між вищими значеннями ІМТ та наявністю діабету (рис. 3.20).

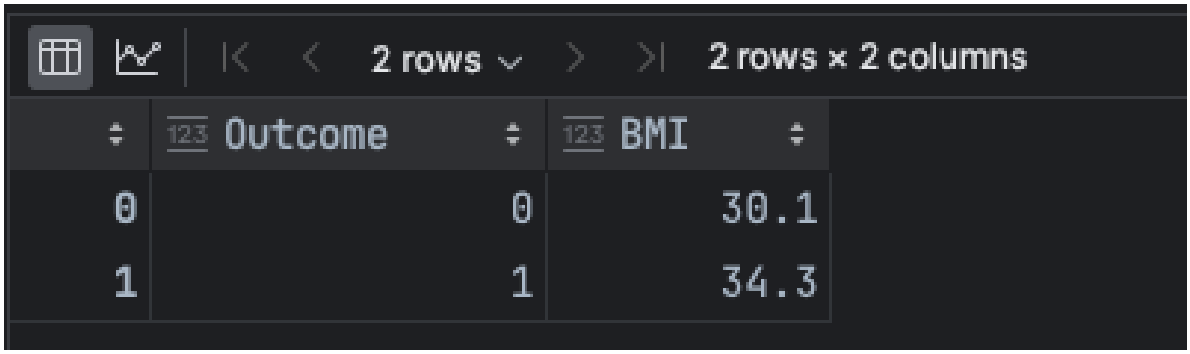

Рисунок 3.20 – Значення індексу маси тіла для здорових та хворих

Розподіл ІМТ в датасеті є асиметричним (skewed), з перекосом у бік вищих значень. Це вказує на те, що в досліджуваній групі присутні особи з високим ІМТ, які можуть значно впливати на середнє значення (рис. 3.21).

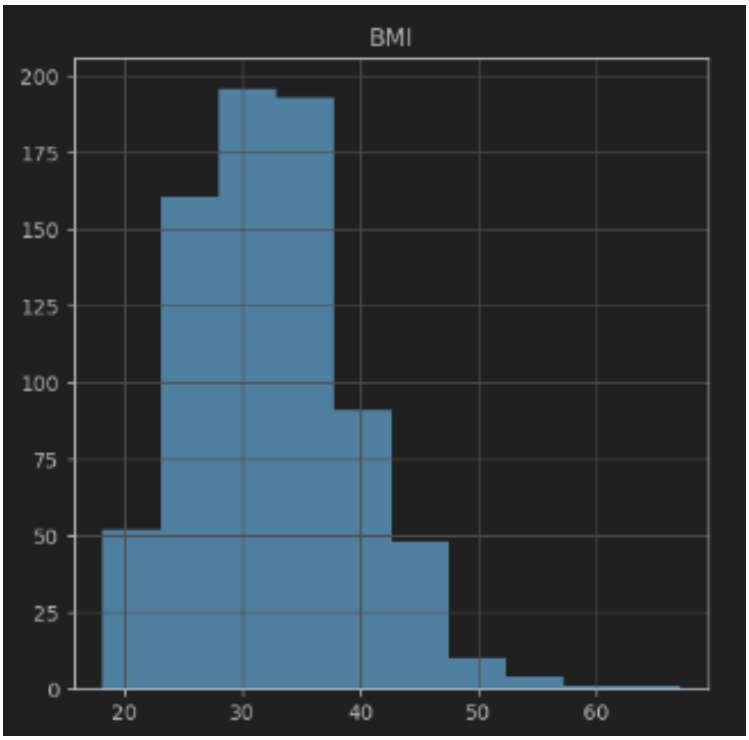

Рисунок 3.21 – Розподілення значень індексу маси тіла

Тому для коректного заповнення відсутніх даних вирішено використовувати медіану, яка краще відображає типове значення ІМТ для цієї вибірки та забезпечує більшу точність аналізу без зміщення, яке могли б спричинити екстремальні значення.

3.3 Аналіз заповнених даних

Після заповнення відсутніх значень важливо провести перевірку заповненості даних для забезпечення їх повноти та готовності до подальшого аналізу та моделювання. Цей крок є критичним, оскільки він дозволяє

підтвердити, що всі очікувані датасети були належним чином оброблені та що в даних більше не залишилося пропусків, які могли б вплинути на результати роботи [29]. Після ретельної перевірки даних було встановлено, що в датасеті відсутні прогалини (рис. 3.22).

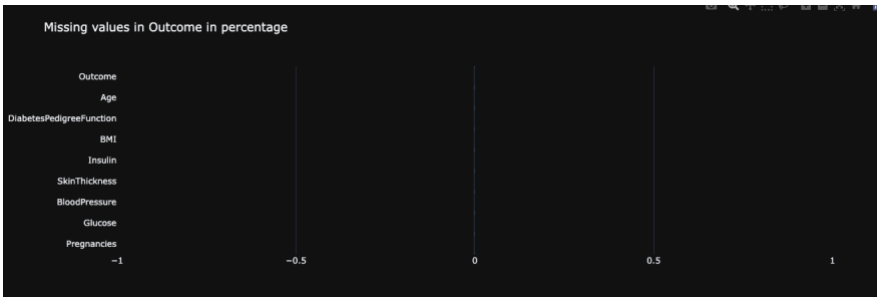

Рисунок 3.22 – Перевірка пропущених значень після їх заповнення

3.4 Тренування моделі та результати

3.4.1 Масштабування даних

Масштабування даних, або scaling, є критично важливим етапом перед тренуванням моделей машинного навчання (рис. 3.23), особливо коли використовується алгоритм *k*-найближчих сусідів (*k*-NN).

Лістинг 3.1 Використання стандартного масштабування у коді

```
sc_X = StandardScaler()
X = pd.DataFrame(sc_X.fit_transform(filled_data.drop(["Outcome"],
axis=1), ),columns=['Pregnancies', 'Glucose', 'BloodPressure',
'SkinThickness', 'Insulin','BMI', 'DiabetesPedigreeFunction', 'Age'])
```

5 rows × 8 columns

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.639947	0.865108	-0.033518	0.670643	-0.181541	0.166619	0.468492	1.425995
1	-0.844885	-1.206162	-0.529859	-0.012301	-0.181541	-0.852200	-0.365061	-0.190672
2	1.233880	2.015813	-0.695306	-0.012301	-0.181541	-1.332500	0.604397	-0.105584
3	-0.844885	-1.074652	-0.529859	-0.695245	-0.540642	-0.633881	-0.920763	-1.041549
4	-1.141852	0.503458	-2.680669	0.670643	0.316566	1.549303	5.484909	-0.020496

Рисунок 3.23 – Результат масштабування моделі

Масштабування обумовлене тим, що *k*-NN визначає схожість між випадками на основі їхніх відстаней в багатовимірному просторі, і наявність атрибутів з різними масштабами може призвести до того, що відстані будуть доміновані тими атрибутами, які мають вищі абсолютні діапазони значень:

– рівномірний вплив: без масштабування, атрибути з великими діапазонами значень можуть непропорційно впливати на визначення відстаней між випадками, що може призвести до спотворення результатів класифікації. Наприклад, атрибут із діапазоном від 0 до 1000 буде впливати на відстань набагато сильніше, ніж атрибут з діапазоном від 0 до 1;

– поліпшення точності: масштабування допомагає забезпечити, що кожен атрибут вносить однаковий вклад у визначення схожості між випадками. Це сприяє більш точному та справедливому обрахунку відстаней, особливо в алгоритмах, які є чутливими до масштабу атрибутів;

– підвищення швидкості навчання: масштабування може також сприяти прискоренню процесів навчання, оскільки оптимізаційні алгоритми часто працюють ефективніше, коли дані знаходяться в однакових масштабах. Це знижує ризик застрягання в локальних мінімумах і сприяє кращій конвергенції алгоритму.

3.4.2 Розділення даних

Розділення даних на тренувальний та тестовий набори є фундаментальним кроком у процесі машинного навчання, оскільки воно дозволяє оцінити якість та ефективність моделі в умовах, що наближені до реального використання [30]. Функція *train_test_split* із бібліотеки sklearn допомагає здійснити це розділення, призначаючи частину даних для навчання моделі, а решту – для її валідації (рис. 3.24).

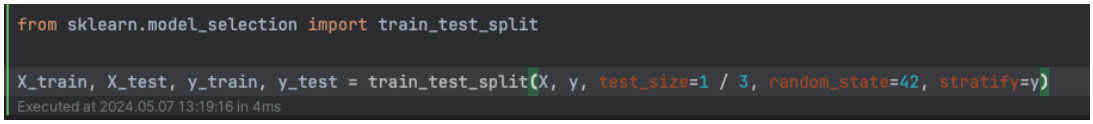

Рисунок 3.24 – Код із розділення даних

В цьому випадку, даних розбиваються таким чином, що одна третина використовується для тестування, а решта – для тренування. Використання параметра *random_state* забезпечує відтворюваність результатів шляхом фіксації початкового стану генератора випадкових чисел, а *stratify=y* гарантує, що у тренувальному та тестовому наборах зберігаються вихідні пропорції класів, представлені в змінній *y*.

3.4.3 Оптимізація параметрів моделі

На початковому етапі було проведено експеримент з різною кількістю сусідів (від 1 до 19), щоб знайти оптимальне значення для *k*-NN.

Лістинг 3.2 Реалізація знаходження оптимального параметра *k*

```
from sklearn.neighbors import KNeighborsClassifier
test_scores = []
train_scores = []
for i in range(1, 20):
    knn = KNeighborsClassifier(i)
    knn.fit(X_train, y_train)
    train_scores.append(knn.score(X_train, y_train))
    test_scores.append(knn.score(X_test, y_test))
```

В результаті аналізу графіка «Train and Test Score with neighbors» найкраще значення параметра *k* виявилося рівним 11, оскільки це значення

забезпечило найвищу збалансованість між точністю на тренувальному та тестовому наборах даних (рис. 3.25).

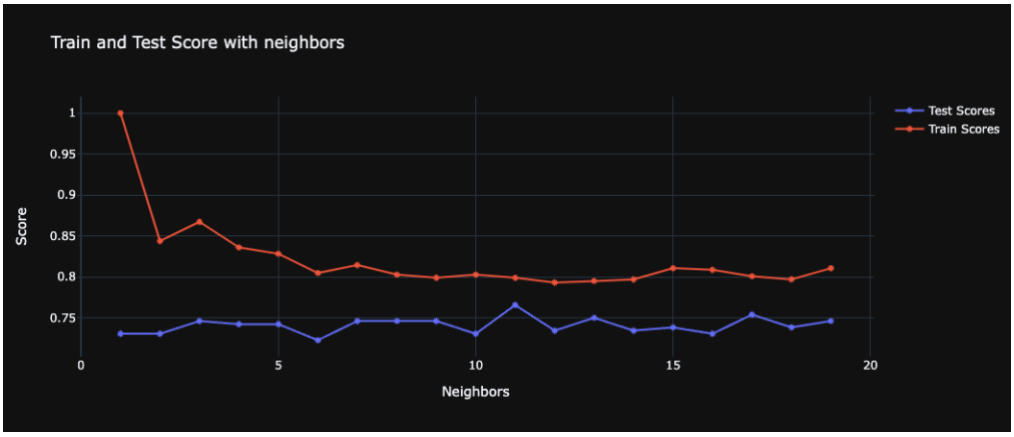

Рисунок 3.25 – Графік «Train and Test Score with neighbors»

3.4.4 Оцінка моделі

Після тренування моделі *k*-NN з використанням 11 сусідів, ефективність моделі була оцінена за допомогою кількох ключових метрик:

– confusion matrix показала наступне: 142 випадки були правильно класифіковані як здорові, а 54 випадки – як хворі, що свідчить про досить високу здатність моделі розрізняти стани (рис. 3.26);

Рисунок 3.26 – Confusion matrix

– звіт про класифікацію вказав на точність 0,77 з фіксованою точністю 0,80 для здорових та 0,68 для хворих, з відповідними показниками відновлення (recall) (рис. 3.27);

	precision	recall	f1-score	support
0	0.80	0.85	0.83	167
1	0.68	0.61	0.64	89
accuracy			0.77	256
macro avg	0.74	0.73	0.73	256
weighted avg	0.76	0.77	0.76	256

Рисунок 3.27 – Звіт про класифікацію

– ROC-крива та AUC-оцінка показали високу здатність моделі відрізняти між класами, з AUC-оцінкою 0,819, що вказує на досить високу загальну ефективність моделі в умовах класифікаційних задач (рис 3.28).

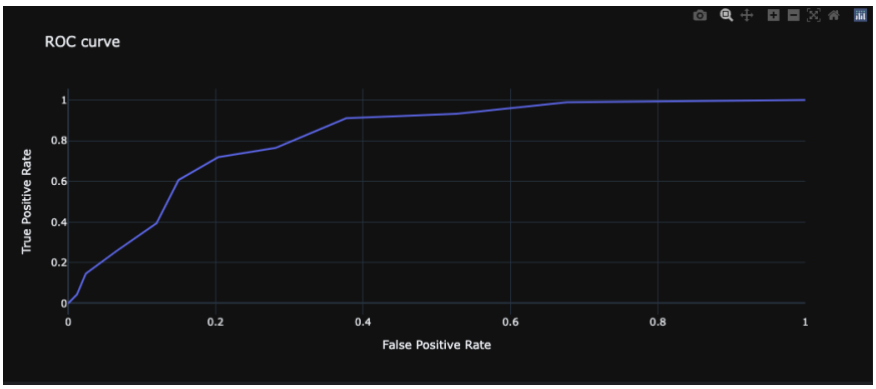

Рисунок 3.28 – ROC крива

3.5 Оптимізація гіперпараметрів

Оптимізація гіперпараметрів є ключовим аспектом у побудові ефективних моделей машинного навчання. Цей процес дозволяє налаштувати модель таким чином, щоб вона досягла найкращої можливої продуктивності.

Для *k*-найближчих сусідів (*k*-NN), основним налаштовуваним гіперпараметром є *n_neighbors*, кількість сусідів, яка безпосередньо впливає на результати класифікації.

Лістинг 3.3 Реалізація оптимізації гіперпараметрів

```
param_grid = {'n_neighbors': np.arange(1, 50)}
knn = KNeighborsClassifier()
knn_cv = GridSearchCV(knn, param_grid, cv=5)
knn_cv.fit(X, y)
```

Після оптимізації результати показали, що найкраща продуктивність моделі досягнута при *n_neighbors* = 25 з точністю приблизно 0.772 (рис. 3.29). Це свідчить, що модель з цим параметром оптимально балансувала між малою та великою кількістю сусідів, забезпечуючи найкращу генералізацію.

```
Best Score:0.7721840251252015
Best Parameters: {'n_neighbors': 25}
```

Рисунок 3.29 – Результат після оптимізації гіперпараметрів за допомогою GridSearchCV

Оптимізація гіперпараметрів через GridSearchCV виявилася вкрай корисною для вдосконалення моделі *k*-NN, ідентифікуючи значення, яке забезпечує найкращу загальну продуктивність. Цей процес є важливим етапом у підготовці моделі до реальних застосувань, оскільки дозволяє адаптувати модель під конкретні умови використання та задачі [31].

Отримані знання можуть бути використані для подальших досліджень та розробки в області прогнозування цукрового діабету, а також для покращення медичних стратегій на основі аналізу даних.

ВИСНОВКИ

У рамках кваліфікаційної роботі було здійснено глибокий аналіз та моделювання для прогнозування розвитку цукрового діабету за допомогою датасету Pima Indians Diabetes. Процес почався з детального розгляду і візуалізації даних, що дозволило виявити ключові особливості та потенційні виклики, як-от відсутні дані та асиметричний розподіл деяких змінних. Наступним кроком було адекватне заповнення відсутніх значень за допомогою медіани чи середнього, залежно від характеру розподілу кожної змінної.

Для підвищення якості даних було застосовано масштабування, що є критично важливим для алгоритмів, заснованих на відстанях, таких як *k*-найближчих сусідів (*k*-NN). В результаті тестування різних значень гіперпараметру *k* за допомогою валідаційного графіка, було обрано оптимальне значення, яке забезпечило найкращий баланс між точністю на тренувальних та тестових наборах.

Застосування моделі *k*-NN з обраним параметром *k* показало хороші результати. Оцінка моделі включала аналіз матриці невідповідностей, звіт про класифікацію та ROC-криву, які всі вказували на адекватну здатність моделі розрізняти між класами. Зокрема, висока значення AUC підтверджує ефективність моделі в контексті обраних метрик.

Додатково, процес оптимізації гіперпараметрів за допомогою GridSearchCV виявив, що подальше підвищення кількості сусідів до 25 може забезпечити ще кращу точність моделі. Це підкреслює важливість тонкого налаштування моделей машинного навчання для конкретних даних та завдань.

Загалом, виконана робота демонструє, як застосування методів машинного навчання може виявити складні закономірності в медичних даних та допомогти в ранньому виявленні захворювань, таких як цукровий діабет. Результати можуть слугувати основою для подальших наукових досліджень та розробки більш точних та ефективних клінічних інструментів.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Global report on Diabetes by WHO (World Health Organization). URL: https://www.who.int/news-room/fact-sheets/detail/diabetes (дата звернення 30.04.2024).

2. Diabetes risk factors. URL: https://www.cdc.gov/diabetes/basics/risk-factors.html (дата звернення 30.04.2024).

3. Diabetes Basics by WHO (World Health Organization). URL: https://www.who.int/health-topics/diabetes (дата звернення 30.04.2024).

4. Тронько, М. Д., Ховака, В.В. & Большова, О. В. (2022) Фармакотерапія ендокринних захворювань. Цукровий діабет та його ускладнення, С. 34-38.

5. Нікберг І.І., & Крайничин Н.Я. (2018) Лікувально-профілактичний режим хворих на цукровий діабет, С. 6-8.

6. Бьюкенен Б. Г., & Шортліфф Е. Х. (1984) Експертні системи, засновані на правилах: Експерименти mycin Стенфордського проекту евристичного програмування, С. 54-55.

7. Тріведі М. К. (2014) Класичний підхід до штучного інтелекту (2-е видання), С. 12-15.

8. A. Agrawal, J. Gans, A. Goldfarb (2019) The economics of artificial intelligence: an agenda, University of Chicago Press, pp. 197-236.

9. T.H. Davenport, W.J. Glover (2018) Artificial intelligence and the augmentation of health care decision-making, pp. 23-25.

10. Y. Guo, Z. Hao, S. Zhao, J. Gong, F. Yang (2020) Artificial intelligence in health care: bibliometric analysis, pp. 3-4.

11. K.-H. Yu, A.L. Beam, I.S. Kohane (2018) Artificial intelligence in healthcare, pp. 719-731.

12. R. Davis (1984) Diagnostic reasoning based on structure and behavior, pp. 34-36.

13. Oleksandra Putyatina, Jörn Sass (2018) Approximation for portfolio optimization in a financial market with shot-noise jumps, pp. 7-8.

14. Iryna Tvoroshenko, Volodymyr Gorokhovatskyi (2022) The Application of Hybrid Intelligence Systems for Dynamic Data Analysis, pp. 21-22.

15. Olga Cherednichenko, Olga Kanishcheva, Olena Yakovleva, Denis Arkatov (2020) Collection and Processing of a Medical Corpus in Ukrainian, pp. 32-33.

16. O. Yakovleva, L. Nebeský, A Kirichenko (2023) Using the GPT models for responses based on custom content to develop neural consultant for university applicants, pp. 31-32.

17. Oleksandr Kuzomin, Vyacheslav Lyashenko (2022) Situational-Linguistic Modeling in Diagnostic Decision-Making Systems, pp. 13-18.

18. Amer Abu-Jassar, Diana Rudenko, Hitham Abdalla (2024) Digital medical image as an object of processing and analysis, pp. 12-13.

19. Oleksandr Kuzomin, Mohammad Ayaz Ahmad, Hryhorii Kots, Vyacheslav Lyashenko, Mariia Tkachenko (2016) Preventing of technogenic risks in the functioning of an industrial enterprise, pp. 77-78.

20. Кузьомін О. Я., Василенко О. О. (2019) Моделювання у процесі проектування інтелектуальної медичної системи діагностування), С. 5-13.

21. Oleksandr Kuzomin, O Dudka, O Vasylenko, V Radchenko, V Lyashenko (2020) Using of ontologies for building databases and knowledge bases for consequences management of emergency, pp. 15.

22. Oleksandr Kuzomin, Oleksandra Dudka, Oleksii Vasylenko, Radion Shylo, Vyacheslav Lyashenko (2020) Mobile Expert System for Diagnostic Human State in Emergency Situations, pp. 19-21.

23. Oleksandr Kuzomin, Oleksandra Dudka, Oleksii Vasylenko, Vyacheslav Lyashenko (2020) The patient organism modeling for diagnosis with the usage of a multi agent representation, pp. 34-35.

24. Oleg Kobylin, Vyacheslav Lyashenko (2020) Time series clustering based on the k-means algorithm, pp. 7-9.

25. Танянський О.С., Руденко Д.О. (2021) Принципи передобробки даних для машинного навчання, С. 20-24.

26. Yevgeniy Bodyanskiy, Sergiy Popov, Filip Brodetskyi, Olha Chala (2022) Adaptive Least-Squares Support Vector Machine and its Combined Learning-Selflearning in Image Recognition Task, pp. 15-17.

27. Y Bodyanskiy, V Skorik, A Deineko, F Brodetskyi (2022) Deep Neural Network with Adaptive Parametric Rectified Linear Units and its Fast Learning, pp. 65-67.

28. Waugh N, Scotland G, McNamee P, Gillett M, Brennan A, Goyder E, Williams R, John (2006) Study Group: Prescreening tools for diabetes and obesity-BMI, waist and waist hip ratio: the D.E.S.I.R, pp. 295-304.

29. Schmidt MI, Duncan BB, Bang H, Pankow JS, Ballantyne CM, Golden SH, Folsom AR, Chambless LE; (2018) The Atherosclerosis Risk in Communities Investigators: Identifying individuals at high risk for diabetes: the Atherosclerosis Risk in Communities study, pp. 324-343.

30. Chen S., Liu L.P., Wang Y.J., Zhou X.H., Dong H., Chen Z.W., Wu J., Gui R., Zhao Q.Y. (2022) Advancing Prediction of Risk of Intraoperative Massive Blood Transfusion in Liver Transplantation with Machine Learning Models. A Multicenter Retrospective Study, pp. 204-208.

31. C. Trocin, P. Mikalef, Z. Papamitsiou, K. Conboy (2023) Responsible AI for digital health: a synthesis and a research agenda, pp. 2123-2157.