Національний лісотехнічний університет України
(повне найменування вищого навчального закладу)
Навчально-науковий інститут комп'ютерних наук та інформаційних технологій
(повне найменування інституту)
Кафедра комп'ютерних наук
(повна назва кафедри)

Магістерська кваліфікаційна робота

другий (магістерський)
(рівень вищої освіти)

на тему: «_Інтелектуальна система прогнозування появи цукрового діабету_»

Виконав: студент _VI_ курсу групи _КН-62м спеціальності
122 "Комп'ютерні науки"
(шифр і назва напряму підготовки, спеціальності)

Василишин Н.Т.
(прізвище та ініціали)
Керівник _Шиманський В.М._____
(прізвище та ініціали)
Рецензент Гентош Л.Ю.
(прізвище та ініціали)

Львів – 2025 року

Національний лісотехнічний університет України
(повне найменування вищого навчального закладу)

ННІ комп`ютерних наук та інформаційних технологій
Кафедра комп`ютерних наук
Рівень вищої освіти другий (магістерський)
Спеціальність 122 «Комп`ютерні науки»
(шифр і назва)

ЗАТВЕРДЖУЮ
Завідувач кафедри КН
_____ Борецька І.Б.
"10" грудня 2025 року

З А В Д А Н Н Я
НА МАГІСТЕРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
Василишину Н.Т.
(прізвище, ім'я, по батькові)

1. Тема роботи *Інтелектуальна система прогнозування появи цукрового діабету*

керівник роботи *Шиманський В.М., к.т.н.*
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від "29" 04 2025 року № С-288
2. Термін подання студентом проекту (роботи) 10.12.2025
3. Вихідні дані до проекту (роботи) Розробити інтелектуальну систему прогнозування появи цукрового діабету. Дослідити швидкодію реалізації алгоритмів навчання. Проаналізувати отримані результати.
4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
Перелік скорочень та умовних позначень. Вступ.
Розділ 1. Стан проблемної області.
Розділ 2. Інформаційне забезпечення
Розділ 3. Математичне забезпечення
Розділ 4. Програмне забезпечення
Розділ 5. Розроблення стартап-проєкту
Висновки. Список використаних джжерел. Додатки

5. Перелік графічного матеріалу (з точним зазначенням обов`язкових креслень)
Слайди для доповіді (підготовка матеріалу для доповіді загальним обсягом 10-12 слайдів)
6. Дата видачі завдання 01.05.2025 року

КАЛЕНДАРНИЙ ПЛАН

№з/п	Назва етапів дипломної роботи	Термін виконання	Відмітка про виконання
1	Аналіз проблематики та визначення напрямків дослідження	01.05.25 - 02.06.25	виконано
2	Аналіз літературних та інформаційних джерел	02.06.25 - 23.06.25	виконано
3	Аналіз даних	23.06.25 - 01.09.25	виконано
4	Попередня обробка даних	01.09.25 - 22.09.25	виконано
5	Вибір та обгрунтування методів та засобів дослідження	22.09.25 - 13.10.25	виконано
6	Реалізація програмного забезпечення.	13.10.25 - 03.11.25	виконано
7	Налаштування інтелектуальної моделі	03.11.25 - 17.11.25	виконано
8	Оформлення пояснювальної записки	17.11.25 - 10.12.25	виконано

Студент: Василишин Н.Т. (підпис)

Керівник роботи: Шиманський В.М. (підпис)

ТЕХНІЧНЕ ЗАВДАННЯ

Проаналізувати існуючі види нейронних меж, що використовуються для вирішення задач класифікації. Вибрати та аргументувати структуру та тип нейронної межі, засобами якої буде проводитись дослідження. Підготувати навчальну вибірку для реалізації нейронної мережі. Розробити інтелектуальну систему прогнозування появи цукрового діабету. Дослідити швидкодію реалізації алгоритмів навчання. Проаналізувати отримані результати.

Розроблена інтелектуальна система повинна задовольняти наступним вимогам:

- мати зручний та інтуїтивно зрозумілий інтерфейс;
- надавати можливість навчати та адаптувати інтелектуальну систему;
- на основі вхідних даних прогнозувати появу цукрового діабету.

АНОТАЦІЯ

Дипломна робота містить 59 сторінок пояснювальної записки, 9 рисунків, 7 таблиць, 1 додаток, 12 джерел.

У роботі представлено розробку інтелектуальної системи прогнозування появи цукрового діабету з використанням сучасних методів машинного навчання та аналізу медичних даних. Актуальність теми зумовлена зростаючою поширеністю діабету у світі та потребою в ефективних засобах ранньої діагностики, що дають змогу своєчасно виявити ризики захворювання та вжити профілактичних заходів.

Основою дослідження став відкритий набір даних BRFSS2015, що містить інформацію про поведінкові, фізіологічні та соціально-демографічні показники респондентів. У роботі виконано повний цикл обробки даних: очищення, нормалізацію, балансування вибірки та відбір ознак. Розглянуто й порівняно декілька моделей машинного навчання, зокрема логістичну регресію, ансамблеві моделі (Random Forest, XGBoost) та нейронні мережі.

Запропонована система дозволяє з високою точністю визначати ймовірність розвитку цукрового діабету на основі індивідуальних характеристик користувача. Проведено оцінювання якості моделей за стандартними метриками (точність, повнота, F-міра, ROC-AUC). Результати підтверджують ефективність застосування штучного інтелекту в медичній сфері, зокрема у задачах прогнозування захворювань.

Розроблена система може стати основою для створення персоналізованих сервісів медичної підтримки, а також для використання у клінічній практиці як допоміжний інструмент лікаря.

Ключові слова: діабет, класифікація, машинне навчання, точність моделі.

ABSTRACT

Thesis contains 59 pages of explanatory note, 9 figures, 7 table, 1 appendix, 12 sources.

The work presents the development of an intelligent system for predicting the onset of diabetes using modern machine learning methods and medical data analysis. The relevance of the topic is due to the growing prevalence of diabetes in the world and the need for effective early diagnostic tools that allow timely detection of disease risks and taking preventive measures.

The research was based on the open BRFSS2015 dataset, which contains information on behavioral, physiological and socio-demographic indicators of respondents. The paper performed a full cycle of data processing: cleaning, normalization, sample balancing and feature selection. Several machine learning models were considered and compared, in particular logistic regression, ensemble models (Random Forest, XGBoost) and neural networks.

The proposed system allows for high-precision determination of the probability of developing diabetes based on individual user characteristics. The quality of the models was assessed using standard metrics (accuracy, completeness, F-measure, ROC-AUC). The results confirm the effectiveness of the application of artificial intelligence in the medical field, in particular in the tasks of disease prediction.

The developed system can become the basis for the creation of personalized medical support services, as well as for use in clinical practice as an auxiliary tool for doctors.

Keywords: diabetes, classification, machine learning, model accuracy.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАЧЕНЬ

AI	Artificial Intelligence — штучний інтелект
ML	Machine Learning — машинне навчання
BRFSS	Behavioral Risk Factor Surveillance System — система нагляду за факторами ризику поведінки
ROC	Receiver Operating Characteristic — крива характеристики роботи приймача
AUC	Area Under Curve — площа під ROC-кривою
TP	True Positive — істинно позитивне спрацювання (правильне виявлення діабету)
TN	True Negative — істинно негативне спрацювання (правильне визначення відсутності діабету)
FP	False Positive — хибнопозитивне спрацювання
FN	False Negative — хибнонегативне спрацювання
F1	F-міра — гармонійне середнє точності й повноти
RF	Random Forest — випадковий ліс
SMOTE	Synthetic Minority Over-sampling Technique — метод синтетичного перенавчання для балансування класів
BMI	Body Mass Index — індекс маси тіла
CSV	

ВСТУП

Актуальність дослідження. За даними ВООЗ, кількість людей, які страждають на цукровий діабет, стрімко зростає. Очікується, що до 2030 року ця хвороба стане однією з головних причин смертності у світі. Найбільш тривожним є той факт, що значна частина хворих не знають про свій діагноз, особливо на ранніх стадіях. Цукровий діабет, особливо типу 2, часто розвивається поступово і може тривалий час не проявлятися явними симптомами. Рання діагностика дозволяє уникнути ускладнень, таких як серцево-судинні захворювання, нефропатії, ретинопатії тощо.

Звичайні методи діагностики базуються переважно на лабораторних аналізах і часто не враховують складну комбінацію факторів ризику, таких як: генетика, спосіб життя, індекс маси тіла, рівень фізичної активності, харчування, тощо. Це відкриває простір для використання штучного інтелекту. Сучасні методи машинного навчання дозволяють аналізувати великі масиви медичних даних і виявляти приховані закономірності. Інтелектуальні системи можуть забезпечити більш точне та персоналізоване прогнозування ризику захворювання, що підвищує ефективність профілактики.

Поява або ускладнення цукрового діабету несе значне економічне навантаження як для системи охорони здоров’я, так і для самих пацієнтів. Запровадження інтелектуальної системи прогнозування допоможе зменшити кількість нових випадків захворювання, оптимізувати ресурси та знизити витрати на лікування.

Об’єкт дослідження – процес раннього виявлення та прогнозування ризику розвитку цукрового діабету.

Предмет дослідження – методи та алгоритми машинного навчання, що застосовуються для аналізу медичних даних і прогнозування появи цукрового діабету.

Метою роботи є розробити інтелектуальну систему прогнозування ризику виникнення цукрового діабету з використанням методів машинного навчання для підвищення ефективності раннього виявлення захворювання та профілактичних заходів..

Для досягнення мети необхідно вирішити ряд **задач**, а саме:

- Провести аналіз існуючих методів прогнозування цукрового діабету та визначити їхні переваги й недоліки.
- Зібрати та підготувати датасет з медичними даними для навчання моделі.
- Обрати та реалізувати алгоритми машинного навчання для побудови моделі прогнозування.
- Провести оцінку точності, чутливості та специфічності моделі.
- Проаналізувати результати роботи системи та порівняти з існуючими підходами.

Наукова новизна дослідження - запропоновано комплексний підхід до прогнозування цукрового діабету на основі медичних та соціально-демографічних даних із використанням алгоритмів штучного інтелекту, що дозволяє досягти вищої точності в порівнянні з традиційними методами.

Практична значимість полягає у тому, що кількість людей з наявністю цукрового діабету зростає, проте це захворювання можна попередити/діагностувати за допомогою класифікації поведінкових факторів ризику.

РОЗДІЛ 1. СТАН ПРОБЛЕМНОЇ ОБЛАСТІ

1.1 Існуючі методи виявлення та прогнозування діабету

Цукровий діабет (ЦД) — це хронічне захворювання, яке виникає через порушення вироблення або дії інсуліну, що призводить до підвищеного рівня глюкози в крові. Згідно зі статистичними даними Всесвітньої організації охорони здоров'я (ВООЗ), на цукровий діабет страждають сотні мільйонів людей у світі, і ця цифра щорічно зростає. Особливо турбує той факт, що значна кількість випадків (особливо типу 2) залишається недіагностованою на ранніх стадіях. Саме тому виявлення та прогнозування цього захворювання є ключовими завданнями у сфері медицини та охорони здоров'я.

Існуючі методи можна умовно поділити на дві основні категорії:

- Традиційні клініко-лабораторні методи діагностики
- Методи оцінки ризику та прогнозування
- Розглянемо кожен із підходів більш детально.

Традиційні методи діагностики ґрунтуються на аналізі біохімічних показників крові та сечі, що дозволяють визначити рівень глюкози, функціональну активність підшлункової залози, а також наявність супутніх порушень.

До основних лабораторних методів належать:

- Вимірювання рівня глюкози в крові натще. Нормальним вважається рівень до 5,5 ммоль/л. Значення понад 7,0 ммоль/л можуть свідчити про наявність цукрового діабету.
- Оральний глюкозотолерантний тест (ОГТТ). Цей тест визначає, як організм реагує на введення глюкози. Його результати є індикатором толерантності до глюкози, що особливо важливо для виявлення переддіабетного стану.

- Глікозильований гемоглобін (HbA1c). Показник, який відображає середній рівень глюкози в крові за останні 2-3 місяці. Значення понад 6,5% свідчить про наявність діабету.
- Визначення цукру в сечі (глюкозурія). Хоча цей метод вважається менш точним, він все ще застосовується для додаткового моніторингу.

Ці методи дозволяють точно встановити наявність захворювання, але мають обмеження в плані прогнозування — вони не завжди дають змогу передбачити ризик розвитку діабету у здорових людей.

У зв'язку з труднощами у ранньому виявленні, з'явились різноманітні методики скринінгу, що дозволяють визначити ризик розвитку цукрового діабету без необхідності в лабораторних дослідженнях. Найпоширенішими є шкали, побудовані на анкетуванні, де кожному фактору ризику надається певна кількість балів.

FINDRISC (Finnish Diabetes Risk Score) — фінська шкала оцінки ризику, що включає питання про вік, ІМТ, окружність талії, фізичну активність, харчові звички, прийом ліків від тиску, рівень глюкози та сімейний анамнез. Вона проста у використанні, не потребує лабораторної діагностики й дозволяє приблизно оцінити 10-річний ризик розвитку діабету 2 типу.

ADA Risk Test (Американська діабетична асоціація) — аналогічний за структурою тест, який на основі базових медико-соціальних параметрів визначає ймовірність наявності або появи діабету.

CANRISK — канадська шкала, адаптована для багатокультурного населення, враховує етнічну приналежність, що є важливим фактором у розвитку діабету.

Перевагами таких систем є їх доступність, швидкість, низька вартість і можливість застосування у великомасштабних епідеміологічних дослідженнях. Водночас їх точність обмежена, оскільки вони не враховують індивідуальні біомаркери, генетику та інші складні взаємозв'язки.

Існують також математичні моделі прогнозування, які базуються на великомасштабних епідеміологічних даних. Такі моделі використовуються для визначення тенденцій поширення захворювання у популяціях. Вони застосовуються на рівні національних та міжнародних організацій охорони здоров'я.

Прикладом є моделі Міжнародної федерації діабету (IDF), які прогнозують кількість хворих на діабет у певних регіонах на основі соціально-демографічних та медичних даних. Проте такі підходи більше орієнтовані на популяційний рівень, а не на індивідуальне прогнозування.

Попри ефективність традиційних підходів у встановленні діагнозу, вони мають низку суттєвих недоліків у контексті прогнозування:

- Висока залежність від моменту вимірювання. Рівень глюкози може змінюватися протягом дня та в залежності від умов (стрес, їжа, фізична активність).
- Неможливість врахувати складні зв'язки між факторами. Наприклад, комбінація певних генетичних і поведінкових факторів може суттєво впливати на ризик, але залишатися непоміченою при використанні класичних шкал.
- Високі витрати на масове тестування. Проведення біохімічних аналізів у великій популяції є дорогим і не завжди доступним.
- Відсутність персоналізованого підходу. Більшість існуючих методів працюють за універсальними шаблонами без урахування індивідуальних характеристик.

У зв'язку з цим, усе більшої популярності набувають інтелектуальні методи аналізу даних, що дозволяють будувати персоналізовані прогностичні моделі.

Сучасні підходи до прогнозування ризику розвитку цукрового діабету все більше базуються на використанні методів машинного навчання та штучного інтелекту (ШІ). Такі моделі здатні аналізувати великі обсяги неоднорідних медичних даних (так звані Big Data) та виявляти приховані

закономірності, які важко ідентифікувати за допомогою класичних статистичних методів.

Алгоритми ШІ використовують класифікацію, кластеризацію, регресійний аналіз та інші підходи для побудови прогнозних моделей. Основна ідея полягає в тому, щоб навчити комп'ютерну модель на основі історичних даних (наприклад, медичних карт пацієнтів) розпізнавати ознаки, що передують розвитку діабету.

Важливим аспектом є також можливість інтеграції таких систем у медичні інформаційні системи, що дозволяє отримувати рекомендації для лікарів або пацієнтів в режимі реального часу.

1.2 Можливості штучного інтелекту у прогнозуванні захворювань

Штучний інтелект (ШІ) останніми роками перетворився з академічної концепції на потужний інструмент, що активно впроваджується в різні сфери, зокрема в медицину. Прогрес в обчислювальній техніці, зростання обсягів медичних даних та розвиток алгоритмів машинного навчання зробили можливим створення систем, здатних самостійно аналізувати великі обсяги інформації та робити прогнози з точністю, що конкурує або навіть перевищує рівень лікарської експертизи.

Застосування ШІ у прогнозуванні захворювань — один із найперспективніших напрямів, особливо в контексті неінфекційних хронічних хвороб, таких як цукровий діабет, серцево-судинні захворювання, онкопатології тощо. Ці стани часто розвиваються поступово, тому можливість раннього прогнозування відіграє вирішальну роль у запобіганні ускладненням і зниженні смертності.

ШІ охоплює кілька технологічних підходів, що застосовуються в аналізі медичних даних.

Машинне навчання (Machine Learning, ML) — підхід, при якому алгоритми навчаються на основі історичних даних і вчаться робити передбачення або класифікації без явного програмування. Найчастіше використовувані алгоритми включають логістичну регресію, дерева рішень,

метод опорних векторів, випадкові ліси (Random Forest), градієнтний бустинг (XGBoost) тощо [2, 8, 9, 11, 12].

Глибинне навчання (Deep Learning, DL) — підвид машинного навчання, який використовує багатошарові нейронні мережі. Глибинне навчання особливо ефективне у задачах розпізнавання образів, мовлення, а також при аналізі неструктурованих даних, наприклад, зображень або текстів медичних записів [6, 7].

Обробка природної мови (Natural Language Processing, NLP) — застосовується для автоматичного аналізу та витягу інформації з неструктурованих медичних текстів, таких як історії хвороб, протоколи обстежень, лікарські висновки тощо.

Зміцнювальне навчання (Reinforcement Learning) — підхід, що дозволяє системі вчитися шляхом отримання зворотного зв’язку за дії. Хоча поки що рідко використовується в клінічному прогнозуванні, має перспективи в адаптивному лікуванні та управлінні хронічними станами [5].

Застосування ШІ дозволяє вирішувати широкий спектр завдань у медицині, включно з:

- Раннє виявлення захворювань на основі симптомів, аналізів, медичних зображень;
- Прогнозування ймовірності розвитку хвороби на основі індивідуальних факторів ризику;
- Оцінка ефективності лікування;
- Створення персоналізованих стратегій профілактики.

ШІ особливо ефективний у тих випадках, де існує багатофакторний вплив, який складно врахувати за допомогою традиційних методів. Наприклад, ризик розвитку цукрового діабету залежить не лише від рівня глюкози в крові, а й від ІМТ, віку, генетичних факторів, рівня фізичної активності, способу життя, психологічного стану тощо.

Машинне навчання активно застосовується для аналізу відкритих датасетів (наприклад, Pima Indians Diabetes Dataset) з метою створення

моделей, які на основі параметрів (вік, ІМТ, кількість вагітностей, рівень глюкози, інсуліну тощо) передбачають ймовірність наявності або розвитку діабету 2 типу [4]. У дослідженнях застосовуються такі алгоритми, як:

- Random Forest, який демонструє високу чутливість і специфічність
- XGBoost, що часто перевершує інші алгоритми за точністю
- Нейронні мережі, здатні моделювати складні нелінійні взаємозв'язки

Інтелектуальні моделі використовуються для прогнозу серцевих нападів або інсультів на основі ЕКГ, артеріального тиску, історії хвороби, віку та інших факторів. Наприклад, дослідження Framingham Heart Study стало основою для численних моделей прогнозування ризику серцево-судинних подій.

Під час пандемії COVID-19 ШІ застосовувався для прогнозування прогресування хвороби, оцінки ризику госпіталізації, виявлення ускладнень за КТ-знімками легень тощо. Це дозволило медичним системам ефективніше розподіляти ресурси [12].

Для ефективного застосування ШІ в медицині важливо інтегрувати аналітичні моделі в електронні медичні системи (ЕМС), клінічні рішення, мобільні додатки. Така інтеграція дозволяє:

- Автоматично виявляти пацієнтів із високим ризиком
- Пропонувати лікарю варіанти діагностики або лікування
- Моніторити ефективність терапії в реальному часі
- Покращувати взаємодію між пацієнтом і медичною установою

Очікується, що в найближчому майбутньому ШІ в медицині буде розвиватися за кількома напрямами:

- Персоналізована медицина — створення індивідуальних моделей ризику для кожного пацієнта
- Мультиоміка — інтеграція генетичних, метаболічних, протеомних даних у прогностичні моделі

- Телемедицина та носимі пристрої — аналіз даних у режимі реального часу (наприклад, глюкометри, фітнес-браслети)
- Самонавчальні системи — адаптація моделей до нових типів даних без ручного втручання

Також велике значення матимуть міждисциплінарні дослідження на стику медицини, комп’ютерних наук та біоінформатики.

1.3. Висновки до розділу

Створення інтелектуальної системи прогнозування появи цукрового діабету є надзвичайно актуальним і перспективним напрямком, який поєднує в собі досягнення медицини, інформатики та аналітики даних. Така система може стати потужним інструментом у профілактиці, ранньому виявленні та ефективному управлінні ризиками розвитку хвороби.

РОЗДІЛ 2. ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ

2.1. Опис даних

У процесі розробки інтелектуальної системи прогнозування появи цукрового діабету надзвичайно важливим елементом є вибір достовірного, репрезентативного і релевантного набору даних. Одним із широко вживаних джерел для навчання і тестування моделей машинного навчання у сфері діабетології є набір даних. Датасет було сформовано на основі щорічного дослідження BRFSS (Behavioral Risk Factor Surveillance System) у 2015 році.

BRFSS (Behavioral Risk Factor Surveillance System) — це програма Центрів з контролю та профілактики захворювань США (CDC), яка щорічно проводить опитування дорослого населення Сполучених Штатів щодо стану здоров'я, поведінкових ризиків, хронічних захворювань та доступу до медичних послуг. BRFSS є одним із найбільших у світі джерел відкритих даних про здоров’я населення, охоплюючи мільйони людей.

У 2015 році зібрані дані включали широкий спектр змінних, пов’язаних із фізичною активністю, вагою, курінням, споживанням алкоголю, доступом до медичної допомоги, соціально-демографічними характеристиками та наявністю захворювань, зокрема діабету.

Набір даних BRFSS2015 є структурованою підмноженою повного BRFSS-опитування, призначеним спеціально для задач бінарної класифікації: наявність або відсутність діабету у респондентів. Датасет оптимізований для використання у машинному навчанні, з фокусом на ознаки, які можуть бути корисними для прогнозування [3, 10].

Кількість записів (рядків): понад 253 000 (тобто понад чверть мільйона унікальних респондентів). Кількість ознак (стовпців) - 22 змінні. Цільова змінна - Diabetes_binary (бінарна: 0 – немає діабету, 1 – є діабет або переддіабет)

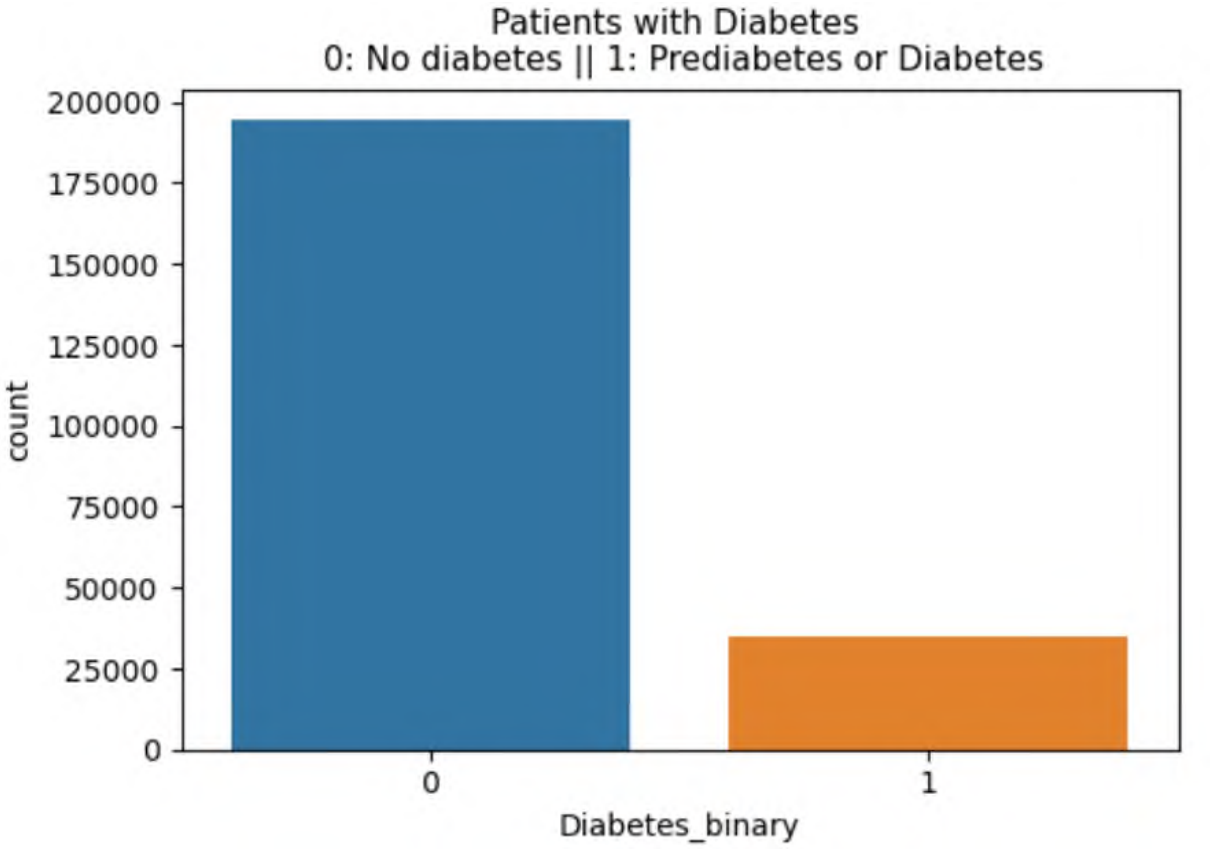

Рисунок 2.1 - Розподіл записів у цільовій ознаці

Такий обсяг даних забезпечує високу надійність та статистичну значущість для побудови моделей. Висока розмірність датасету дозволяє моделі краще узагальнювати залежності. Клас «без діабету» переважає, що потребує балансування вибірки. Персональні ідентифікатори видалені, що відповідає вимогам конфіденційності.

Датасет ідеально підходить для побудови моделей бінарної класифікації. Його можна використовувати для:

- Аналізу факторів ризику розвитку діабету;
- Побудови інтелектуальних систем раннього попередження;
- Навчання моделей штучного інтелекту (SVM, Random Forest, XGBoost, Neural Networks тощо);
- Оцінки впливу поведінкових і соціально-економічних факторів на здоров’я.

Попри переваги, датасет має певні обмеження:

- Дані зібрані лише у США, що може вплинути на узагальнення для інших країн;
- Самозвітність (self-reported data) – суб’єктивні відповіді можуть бути неточними;
- Дані не включають клінічні лабораторні тести, які є критичними у деяких випадках;
- Категоризація віку та доходу — інтервальна, а не точна.

Набір даних BRFSS2015 є цінним ресурсом для розробки та тестування моделей прогнозування ризику розвитку цукрового діабету. Його велика кількість записів, зручна структура, наявність ключових змінних та фокус на бінарну класифікацію роблять його ідеальним для застосування у системах штучного інтелекту, спрямованих на покращення медичної діагностики та профілактики. Водночас правильне розуміння обмежень і якісна попередня обробка даних є критичними для досягнення високої точності й надійності побудованих моделей.

2.2. Процедури попереднього опрацювання даних

Процедури попереднього опрацювання даних є одним з найважливіших етапів при створенні інтелектуальної системи прогнозування появи цукрового діабету. У практичних задачах медичної інформатики дані часто є неповними, шумними, неоднорідними або несумісними за структурою. Наявність таких недоліків може суттєво знизити ефективність навіть найсучасніших математичних моделей. Тому перед застосуванням алгоритмів машинного навчання або статистичного аналізу необхідно провести детальну очистку, нормалізацію, трансформацію та фільтрацію даних.

Першим кроком є аналіз структури та змісту вхідних даних. Це включає:

- Типи змінних у медичних датасетах, що використовуються для прогнозування діабету, можуть бути присутні:

- Числові ознаки (наприклад, рівень глюкози в крові, вік, індекс маси тіла, артеріальний тиск);
- Категоріальні ознаки (наприклад, стать, наявність спадкової схильності, етнічна група);
- Бінарні змінні (наприклад, курить/не курить);
- Часові ряди (якщо йдеться про динамічне відстеження параметрів).

Важливо перевірити, чи представлені дані у вигляді таблиці, з рядками, що відповідають окремим пацієнтам або спостереженням, і стовпцями, що містять змінні.

У медичних даних пропущені значення — явище досить поширене. Причинами можуть бути технічні збої, людський фактор або специфіка збору даних (не всі пацієнти проходять однакові тести). Розрізняють кілька стратегій обробки таких значень. Видалення записів – застосовується, коли кількість пропусків невелика, і це не впливає на статистичну репрезентативність. Недоліком є втрата потенційно корисної інформації. Заповнення середнім/медіаною/модою – для числових змінних можливо використати середнє значення, для категоріальних – найчастіше значення. Інтерполяція – для часових рядів використовується лінійна або поліноміальна інтерполяція. Передбачення пропущених значень – застосовуються моделі (наприклад, k-найближчих сусідів), які прогнозують значення змінної на основі інших наявних параметрів. Позначення відсутності як окреме значення – підходить для категоріальних даних, де «невідомо» може мати власне значення.

Аномальні або викидні значення можуть спотворити результати моделювання, особливо у медичних задачах. Для обробки викидів їх або видаляють, або замінюють на межові значення, або застосовують трансформацію (логарифмування, нормалізація).

Більшість алгоритмів машинного навчання вимагає однакового масштабу змінних. Наприклад, глюкоза може варіюватися від 50 до 300

мг/дл, тоді як вік — від 0 до 100 років. Якщо не провести масштабування, такі дисбаланси можуть знецінити вагу окремих параметрів у моделі.

Багато моделей не можуть працювати з текстовими змінними. Тому категоріальні змінні слід трансформувати у числові. Label Encoding – присвоює кожному унікальному значенню числовий код (наприклад, «чоловік» = 0, «жінка» = 1). One-Hot Encoding – створює новий бінарний стовпець для кожного значення категорії. Наприклад, «етнічна група» → [білий, чорний, азіат]. Target Encoding – категоріальні значення замінюються на середнє значення цільової змінної в цій групі (ризиковано, бо може призводити до переобучення).

У більшості відкритих датасетів щодо діабету клас «здоровий» зустрічається частіше, ніж клас «хворий». Такий дисбаланс може викликати упередженість моделі до домінуючого класу.

Oversampling – штучне дублювання зразків меншого класу (наприклад, SMOTE – Synthetic Minority Over-sampling Technique). Undersampling – зменшення кількості записів домінуючого класу. Комбіновані підходи – балансування за допомогою гібридних технік.

Для оцінювання моделі важливо мати незалежну тестову вибірку. Зазвичай дані діляться у співвідношенні 70/30 або 80/20. У складніших задачах використовують крос-валідацію (k-fold cross-validation) або стратифіковану вибірку — щоб забезпечити однаковий розподіл класів у тренувальній і тестовій вибірках.

Попереднє опрацювання даних — це не просто технічний етап, а фундамент для побудови надійної та точнішої інтелектуальної системи прогнозування цукрового діабету.

2.3. Висновки до розділу

Від якості підготовки даних залежить кінцева продуктивність моделей машинного навчання, рівень інтерпретованості результатів та здатність

системи до узагальнення. У сфері охорони здоров’я, де йдеться про людське життя, цей етап набуває критичного значення. Систематичне застосування вищезазначених процедур дозволяє створити достовірну, об'єктивну та ефективну модель, що здатна виявляти ризики розвитку діабету на ранніх стадіях.

РОЗДІЛ 3. МАТЕМАТИЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Огляд математичних методів прогнозування ризику діабету

Цукровий діабет (ЦД), зокрема другого типу, є одним із найпоширеніших хронічних захворювань у світі, що пов'язане з порушенням обміну глюкози. Через свою прогресуючу природу, здатність до безсимптомного перебігу на початкових етапах та високий ризик ускладнень, своєчасне прогнозування й виявлення цього захворювання є критично важливим. У зв’язку з цим застосування математичних та статистичних методів прогнозування, зокрема методів штучного інтелекту, є одним із найбільш перспективних напрямів у медичній інформатиці.

Розглянемо ключові математичні методи, які можуть бути використані для задачі прогнозування ризику цукрового діабету. Ці методи умовно можна поділити на три основні групи: класичні статистичні методи; машинне навчання; нейронні мережі та глибинне навчання [2, 9, 11].

3.1.1. Логістична регресія

Логістична регресія — один із найбільш використовуваних методів бінарної класифікації в медицині. Вона дозволяє змоделювати ймовірність появи цукрового діабету на основі незалежних змінних (наприклад, вік, індекс маси тіла, рівень глюкози, наявність спадкових факторів).

Формально модель має вигляд:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}} \tag{3.1}$$

Переваги методу:

- Інтерпретованість (коефіцієнти мають статистичний зміст).
- Висока швидкість обчислень.
- Відносна простота реалізації.

Недоліки:

- Лінійність — модель не вловлює складні взаємозв’язки між змінними.
- Чутливість до мультиколінеарності та викидів.

3.1.2. Дерева рішень

Метод дерев рішень (Decision Trees) — простий і зрозумілий підхід до класифікації, що базується на ітеративному поділі простору характеристик. Рішення у вузлах приймаються за критерієм максимальної інформаційної вигоди або мінімізації ентропії.

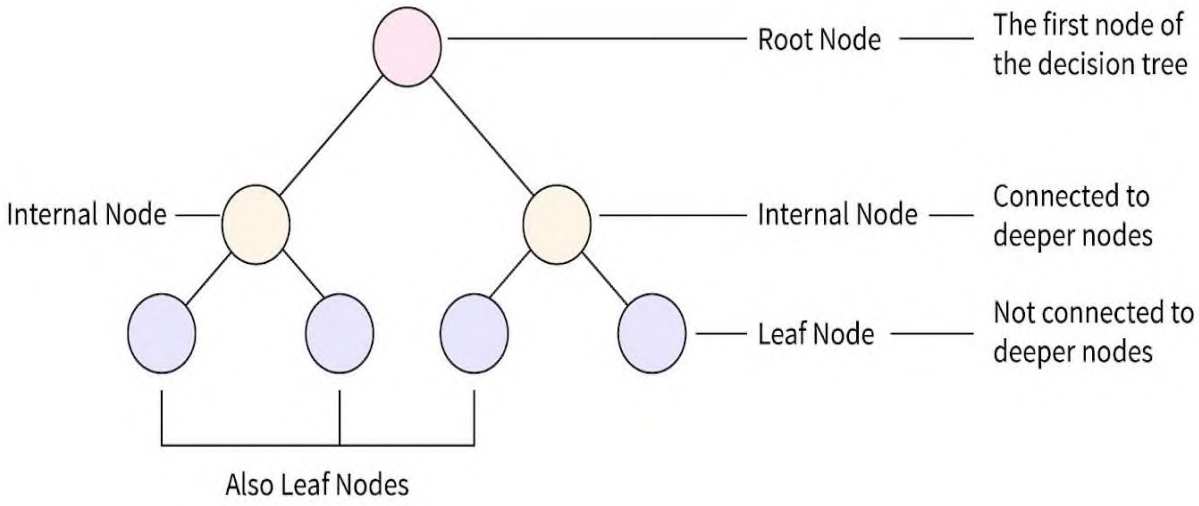

Рисунок 3.1 - Приклад дерева рішень

Переваги:

- Добра інтерпретованість.
- Можливість роботи з нечисловими змінними.
- Виявлення важливих ознак.

Недоліки:

- Низька стабільність (малозначні зміни в даних можуть сильно змінити дерево).
- Схильність до перенавчання.

3.1.3. Випадковий ліс (Random Forest)

Random Forest (Випадковий ліс) — це ансамблевий алгоритм машинного навчання, який поєднує велику кількість дерев рішень (decision trees) для отримання більш стабільного, точного та стійкого до перенавчання прогнозу.

Цей метод можна використовувати для:

- Класифікації
- Регресії
- Оцінки важливості ознак (feature importance)

Random Forest працює за принципом ансамблю дерев рішень, де кожне дерево навчається на різних підвибірках даних і ознак. Потім його рішення об’єднуються. Для класифікації — голосування більшості (majority voting). Для регресії — середнє арифметичне прогнозів. Це дозволяє зменшити дисперсію моделі, зменшити ризик перенавчання та покращити узагальнення.

Random Forest працює за наступною схемою:

- Крок 1: Бутстрепінг (bagging). З початкової вибірки створюється кілька випадкових підвибірок із поверненням (bootstrapped samples). Кожне дерево навчається на своїй підвибірці.
- Крок 2: Випадковий вибір ознак. Під час побудови кожного дерева, на кожному розгалуженні (split), розглядається випадкова підмножина ознак, а не всі ознаки. Це створює більше різноманіття між деревами та зменшує кореляцію між ними.
- Крок 3: Комбінування результатів. Для класифікації: кожне дерево "голосує", і обирається клас, за який проголосувала більшість дерев. Для регресії, обчислюється середнє всіх прогнозів дерев.

Навіть якщо окремі дерева дають шумні або слабкі передбачення, загальний ансамбль зазвичай дуже точний. Багато дерев із випадковим

вибором ознак і бутстрепінгом допомагають знизити ризик overfitting. Random Forest може оцінювати внесок кожної ознаки у точність моделі.

Таблиця 3.1. Основні параметри Random Forest

Параметр	Опис
n_estimators	Кількість дерев у лісі
max_depth	Максимальна глибина кожного дерева
max_features	Кількість ознак, що розглядаються на кожному поділі
min_samples_split	Мінімальна кількість зразків для поділу вузла
min_samples_leaf	Мінімальна кількість зразків у листі
Bootstrap	Чи використовувати бутстрепінг
Criterion	Міра якості поділу (gini, entropy, mse, mae)

Random Forest — потужний і надійний алгоритм, що забезпечує високу точність без потреби у надмірному налаштуванні гіперпараметрів. Це чудовий вибір для початку при роботі з табличними даними, особливо коли потрібно зменшити перенавчання або отримати попередню оцінку важливості ознак.

3.1.4. Градієнтний бустинг XGBoost, LightGBM

XGBoost (Extreme Gradient Boosting) — це високоефективна реалізація алгоритму градієнтного бустингу на деревах рішень (GBDT). Вона спеціально оптимізована для: швидкості виконання; ефективного використання пам’яті; високої точності прогнозів

Цей алгоритм часто використовується для задач класифікації, регресії, ранжування та в машинному навчанні для задач з табличними даними. XGBoost широко застосовується у практичних задачах бізнесу, медицини, фінансів, біоінформатики тощо.

Градієнтний бустинг — це ансамблевий метод, що комбінує багато слабких моделей (зазвичай дерев рішень), кожна з яких намагається виправити помилки попередніх.

Ідея полягає в тому, що кожне наступне дерево навчається на залишках помилок попередніх дерев. Це ітеративний процес, кожен крок зменшує функцію втрат (наприклад, логарифмічну втрату або MSE).

XGBoost вводить регуляризацію прямо у функцію втрат, що допомагає боротися з перенавчанням. Функція втрат визначається як

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.2}$$

де

- l — функція втрат (наприклад, квадратична втрата),
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ — регуляризація складності дерева.

XGBoost самостійно визначає оптимальний шлях для missing значень під час навчання. Навчання можна зупинити, якщо протягом N ітерацій не покращується метрика на валідаційній вибірці.

Таблиця 3.2 - Основні параметри XGBoost

Параметр	Опис
n_estimators	Кількість дерев
max_depth	Максимальна глибина дерева
learning_rate (eta)	Коефіцієнт навчання (step size shrinkage)
subsample	Частка зразків, що використовуються для навчання кожного дерева
colsample_bytree	Частка ознак, що використовуються в кожному дереві
Gamma	Мінімальний приріст приросту, необхідний для поділу
lambda, alpha	L2 та L1 регуляризація відповідно

XGBoost — це потужний і гнучкий інструмент для задач машинного навчання, особливо для структурованих табличних даних. Його ефективність, масштабованість і точність зробили його золотим стандартом у багатьох прикладних задачах.

Таблиця 3.3 - Порівняння методів

Метод	**Точність**	**Інтерпретованість**	**Стійкість**	**Час обробки**
Логістична регресія	Середня	Висока	Висока	Швидкий
Random Forest	Висока	Середня	Висока	Середній
XGBoost	Дуже висока	Низька	Висока	Повільний

Усі розглянуті математичні методи мають свої переваги та недоліки, і вибір оптимального підходу залежить від типу даних, вимог до точності, швидкості й інтерпретованості. У випадку прогнозування ризику цукрового діабету особливо важливо зберігати баланс між точністю моделі та її пояснюваністю, адже рішення, що приймаються на її основі, можуть безпосередньо впливати на здоров’я пацієнтів.

3.2 Методи оцінювання якості математичних моделей

Ефективність будь-якої математичної моделі, зокрема моделі прогнозування ризику цукрового діабету, має оцінюватися на основі чітких, об’єктивних та стандартизованих критеріїв. Якість моделі визначає її практичну придатність, точність діагностики, здатність узагальнювати знання на нові дані та мінімізувати кількість помилкових висновків, що є особливо важливим у медичних системах.

Цей підрозділ розкриває основні методи оцінювання якості математичних моделей у задачах класифікації, що широко використовуються

в медичній інформатиці, а також надає приклади їх інтерпретації в контексті прогнозування діабету.

У задачі бінарної класифікації, яку представляє прогнозування наявності або відсутності ризику цукрового діабету, кожен прогноз може бути або правильним, або помилковим. Залежно від істинного класу та результату передбачення моделі, усі рішення можна класифікувати в чотири категорії:

- TP (True Positive) – істинно позитивні: модель правильно передбачила наявність діабету.
- TN (True Negative) – істинно негативні: модель правильно передбачила його відсутність.
- FP (False Positive) – хибно позитивні: модель передбачила діабет, хоча його немає.
- FN (False Negative) – хибно негативні: модель не виявила діабет, хоча він є.

Знаючи ці значення, обчислюють численні метрики якості класифікації.

Метрика Accuracy (точність класифікації)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

Ця метрика показує загальну частку правильно класифікованих випадків. Вона є простою та інтуїтивно зрозумілою, проте може вводити в оману при дисбалансі класів. Наприклад, якщо 90% пацієнтів не мають діабету, модель, що завжди передбачає "немає діабету", матиме 90% точності, хоча зовсім не виявляє хворих.

Метрика Precision (точність позитивного прогнозу)

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

Precision вказує, яка частина передбачених позитивних результатів є правильними. У медичних задачах це важливо, щоб не "помилково налякати" пацієнтів або не призначити непотрібне лікування.

Метрика Recall (повнота або чутливість)

$$Recall = \frac{TP}{TP + FN} \tag{3.5}$$

Це метрика, що показує здатність моделі виявляти хворих. Вона особливо важлива в задачах діагностики, де пропущений випадок діабету (FN) може мати серйозні наслідки.

Метрика F1-score

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.6}$$

F1-score є гармонічним середнім між precision та recall і використовується для балансування між ними. Це особливо корисно, коли потрібно мінімізувати і FP, і FN, що актуально для виявлення цукрового діабету.

Матриця помилок (confusion matrix). Матриця помилок є зручним способом візуального представлення результатів класифікації. Вона дозволяє чітко побачити кількість кожного типу помилок. За цією матрицею легко обчислювати усі вищенаведені метрики.

ROC (Receiver Operating Characteristic) – це графік, який показує співвідношення між чутливістю (True Positive Rate) та специфічністю (False Positive Rate = FP / (FP + TN)) при зміні порогу класифікації.

AUC показує площу під ROC-кривою. Значення AUC:

- – ідеальна модель.
- 0.9–0.99 – дуже гарна модель.
- 0.8–0.9 – прийнятна модель.
- 0.5 – випадкове вгадування.

У задачі прогнозування діабету AUC > 0.85 зазвичай вважається дуже хорошим результатом.

Оцінка якості моделі не обмежується обчисленням метрик — важливо ще й перевірити її здатність до узагальнення. Найбільш простий підхід –

розділення даних на тренувальну (наприклад, 70%) та тестову (30%) частини. На тренуванні модель навчається, на тесті – оцінюється її якість.

Перехресна валідація (k-fold cross-validation), цей метод полягає в розбитті вибірки на k підгруп, де кожна з них по черзі виступає тестовою. Це дозволяє отримати стабільні оцінки якості та уникнути залежності від одного випадкового поділу.

Хибно негативні помилки (FN) вважаються найнебезпечнішими, оскільки призводять до недіагностованих випадків діабету. Відповідно, під час вибору моделі необхідно надавати пріоритет recall.

Хибно позитивні (FP) призводять до непотрібного хвилювання та можливого перевантаження системи охорони здоров'я. У таких випадках важливо контролювати precision.

Моделі з високим AUC та F1-score є оптимальними у випадках, коли існує баланс між виявленням і обережністю.

Якісна оцінка математичної моделі в задачах прогнозування ризику цукрового діабету є критично важливою для забезпечення надійності, безпеки та ефективності прийняття клінічних рішень. Однієї метрики недостатньо — лише в комбінації (наприклад, F1-score, AUC, recall, precision) можна отримати цілісну картину. Крім того, використання методів перехресної валідації дозволяє гарантувати, що модель не тільки гарно працює на тестових даних, але й здатна до узагальнення.

3.3. Висновки до розділу

Для практичної реалізації інтелектуальної системи найдоцільніше розглядати комбінацію методів, зокрема: логістичну регресію — як базову модель, Random Forest або XGBoost — як потужні ансамблеві підходи, а також MLP — для випадків, де доступно більше даних та потрібна висока точність. У підсумку, ретельна оцінка моделей за метриками (точність, повнота, F1-score) дозволяє зробити обґрунтований вибір математичного ядра системи.

РОЗДІЛ 4. ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ

4.1 Аналіз датасету BRFSS

Змістовне розуміння структури, властивостей та статистичних характеристик набору даних є критично важливим для формування ефективної моделі машинного навчання, що дозволяє з високою точністю прогнозувати наявність або ймовірність виникнення цукрового діабету.

BRFSS — найбільша у світі система щорічного телефонного опитування дорослого населення, започаткована Центрами з контролю та профілактики захворювань США (CDC). Вона дозволяє відстежувати поведінкові ризики, пов’язані зі здоров’ям, наявність хронічних хвороб, доступ до медичних послуг тощо. У 2015 році BRFSS охопила мільйони респондентів з усіх штатів США. З усього обсягу даних була сформована спрощена та структурована версія, призначена для завдань класифікації BRFSS2015.

Цей набір даних містить понад 250 000 записів, кожен з яких представляє унікального респондента. Усі значення є числовими (бінарними або категоріальними), що полегшує обробку та уможливлює застосування широкого спектра моделей машинного навчання без необхідності складного кодування ознак.

Датасет містить 22 ознаки, серед яких одна — цільова змінна (Diabetes_binary), а решта — незалежні змінні, які описують поведінкові, фізіологічні, соціально-демографічні та медичні характеристики респондентів.

Цільова змінна має значення:

- 0 — відсутність діабету;
- 1 — діабет або переддіабет.

Найважливіші незалежні змінні:

- HighBP — наявність високого кров'яного тиску;
- HighChol — підвищений рівень холестерину;

- BMI — індекс маси тіла;
- Smoker — факт куріння в минулому;
- PhysActivity — наявність фізичної активності;
- Fruits / Veggies — споживання фруктів та овочів;
- GenHlth — суб’єктивна оцінка здоров’я;
- Income — рівень доходу;
- Education — рівень освіти;
- Age — вікова група;
- інші змінні, що характеризують стан психічного та фізичного здоров’я.

Було виконано обчислення коефіцієнтів кореляції між усіма ознаками та цільовою змінною. Основні спостереження зображені на рис. 4.1.

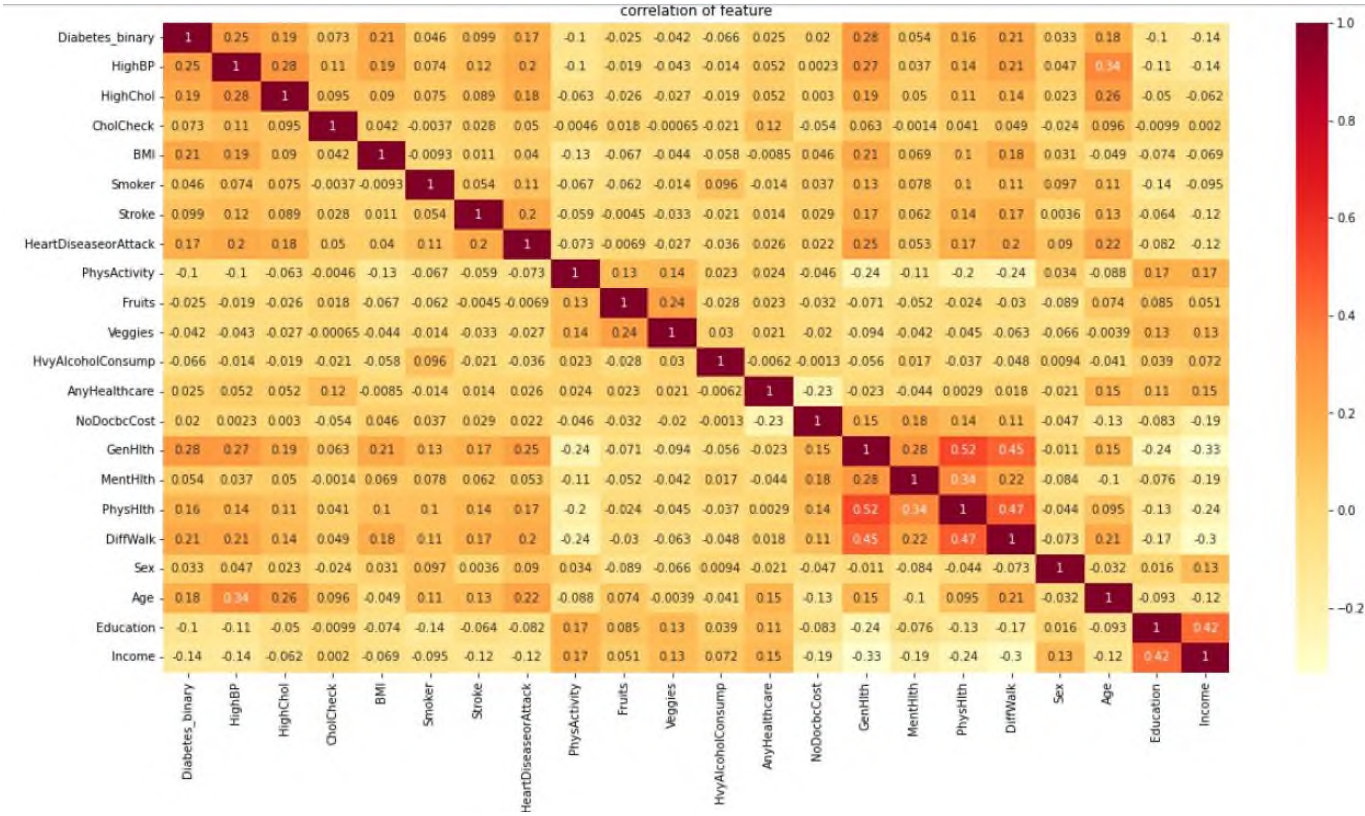

Рисунок 4.1 - Матриця кореляцій

BMI має позитивну кореляцію з діабетом. HighBP і HighChol також демонструють суттєву залежність. PhysActivity — обернено пропорційна

кореляція з діабетом. Fruits і Veggies — слабка, але стабільна негативна кореляція. Це дозволяє стверджувати, що фактори способу життя дійсно мають суттєвий вплив на ризик розвитку діабету.

Було побудовано декілька візуалізацій.

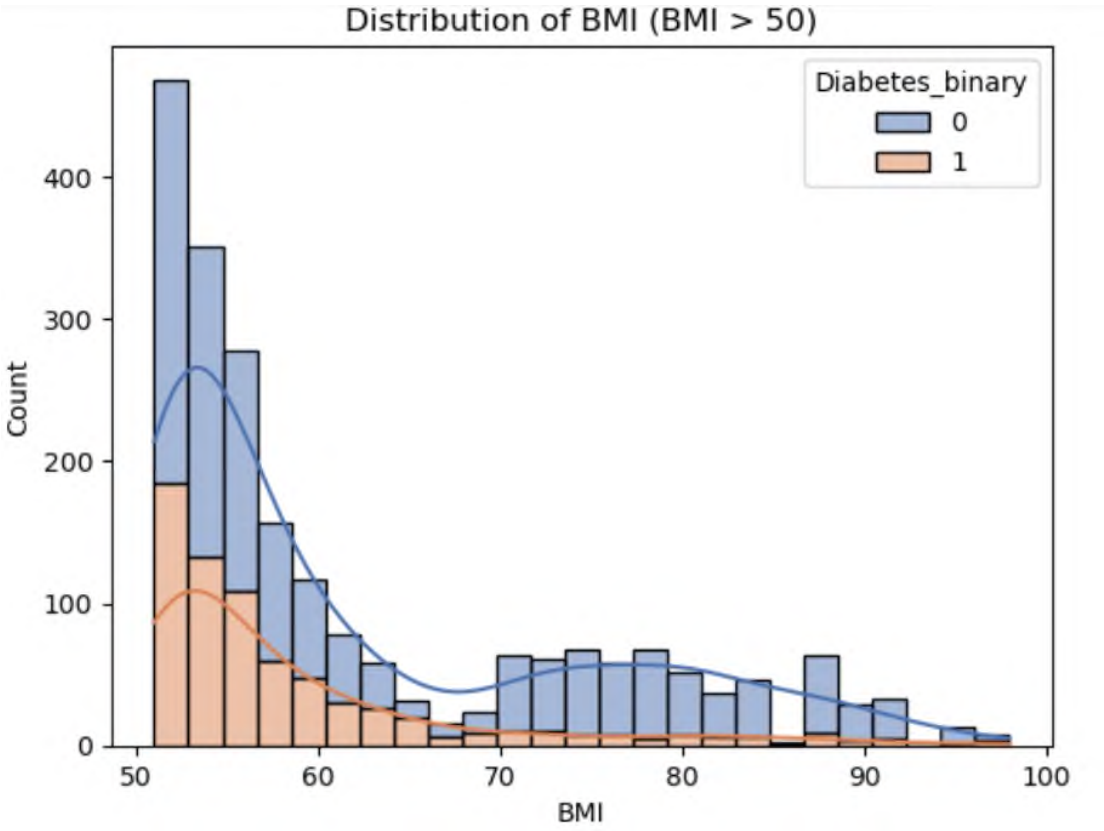

Рисунок 4.2 - Кількість випадків діабету залежно від BMI

Гістограми BMI для класів 0 і 1: у групі з діабетом пік зсунутий вправо (більша вага). Існують чіткі закономірності між віком та наявністю діабету;

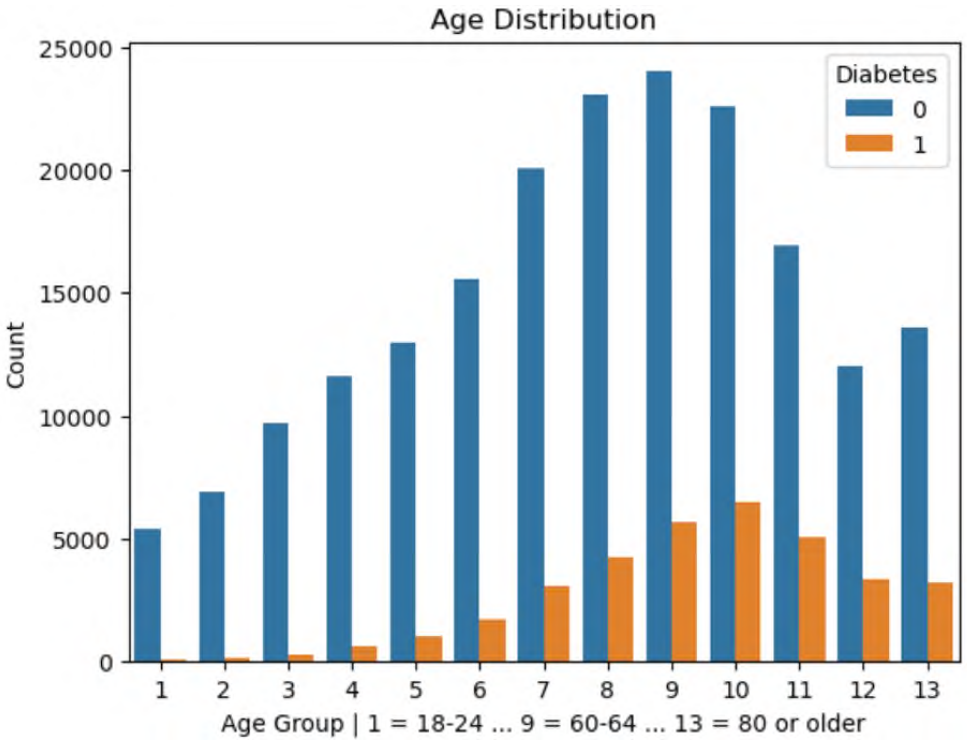

Рисунок 4.3 - Кількість випадків діабету залежно від вікової групи

Дані є репрезентативними для дорослого населення США, але обмежені для екстраполяції на інші регіони без додаткової адаптації.

Аналіз BRFSS-похідного датасету дав змогу виявити найважливіші предиктори ризику діабету та окреслити підходи до їх обробки. Особливості змінних, якість даних та обсяг вибірки забезпечують надійну основу для побудови високоточних моделей прогнозування. З урахуванням виявлених закономірностей та проблем (зокрема дисбалансу), наступні етапи розробки зосереджуються на виборі оптимальної моделі та її навчанні з використанням підготовлених даних.

4.2 Аналіз отриманих результатів

4.2.1 Модель логістичної регресії

Після навчання логістичної регресії на тренувальній вибірці із застосуванням балансування класів, були отримані такі результати на тестовій вибірці [1].

Таблиця 4.1 - Результати навчання логістичної регресії.

Метрика	Значення
Accuracy	0.767
Precision	0.624
Recall	0.582
F1-score	0.602
ROC-AUC	0.832

Отримані результати навчання логістичної регресії після застосування балансування класів свідчать про достатньо стабільну роботу моделі на тестовій вибірці. Значення Accuracy становить 0.767, що означає правильну класифікацію приблизно 77% випадків. Це є непоганим показником для задачі прогнозування ризику захворювання, особливо з урахуванням того, що початковий розподіл класів був нерівномірним, а балансування дозволило уникнути переважання одного класу над іншим. Проте варто зазначити, що точність сама по собі не є достатньою метрикою для оцінки якості моделі в умовах потенційного дисбалансу класів, тому необхідно розглядати інші показники.

Метрика Precision дорівнює 0.624, що свідчить про те, що серед усіх випадків, які модель класифікувала як позитивні (наявність ризику діабету), близько 62% є правильними. Це важливо для медичних задач, адже висока точність зменшує кількість хибнопозитивних результатів, що може знизити непотрібні витрати на додаткові обстеження. Водночас Recall становить 0.582, що означає здатність моделі виявляти реальні позитивні випадки на рівні 58%. Цей показник є критичним у контексті охорони здоров’я, оскільки пропуск пацієнтів із високим ризиком може призвести до серйозних наслідків. Баланс між Precision і Recall відображає значення F1-score, яке дорівнює 0.602. Це свідчить про помірну узгодженість між здатністю моделі правильно ідентифікувати позитивні випадки та уникати помилкових спрацювань.

Особливо важливим є показник ROC-AUC, який становить 0.832. Це досить високий результат, що демонструє хорошу здатність моделі розрізняти класи незалежно від вибраного порогу. Значення понад 0.8 зазвичай вважається добрим рівнем для медичних прогнозів, що підтверджує ефективність використання логістичної регресії в поєднанні з балансуванням класів. Високий ROC-AUC свідчить про те, що модель має потенціал для налаштування порогів класифікації залежно від пріоритетів — наприклад, можна збільшити Recall для зменшення пропусків пацієнтів із ризиком, навіть якщо це трохи знизить Precision.

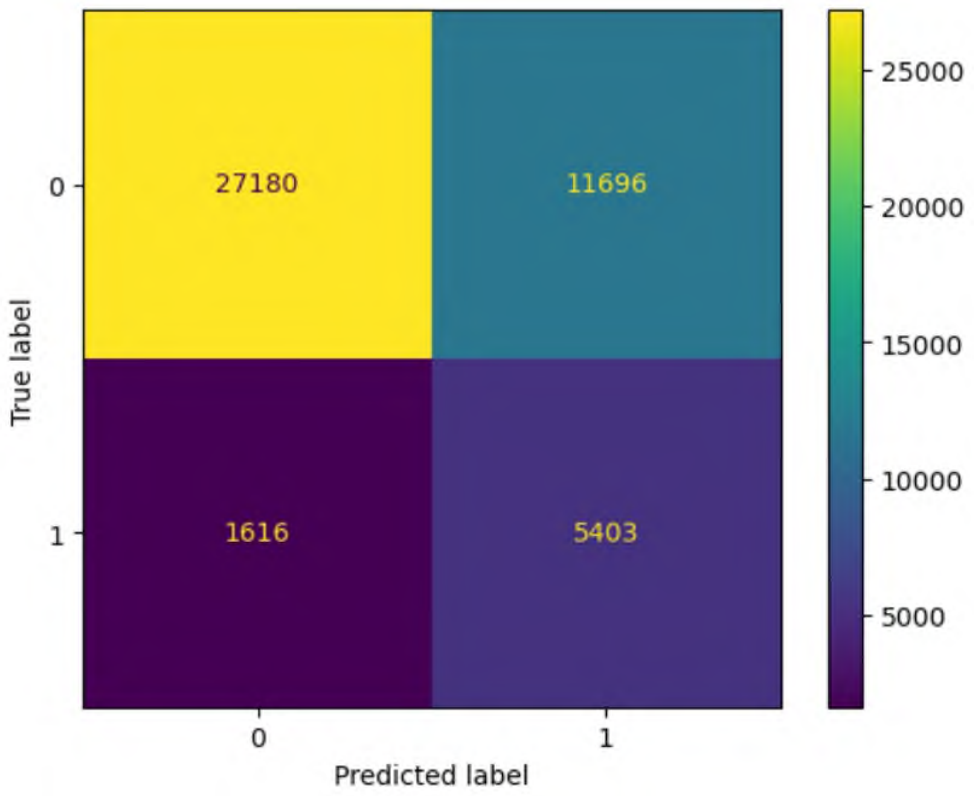

Рисунок 4.4 - Confusion матриця для логістичної регресії

4.2.2 Модель Random Forest

Після оптимізації гіперпараметрів (кількість дерев, максимальна глибина, кількість ознак для розбиття) модель показала наступні результати

Таблиця 4.2 Результати навчання Random Forest.

Метрика	Значення
Accuracy	0.842
Precision	0.715
Recall	0.689
F1-score	0.701
ROC-AUC	0.897

Отримані результати навчання моделі Random Forest після оптимізації гіперпараметрів демонструють суттєве покращення порівняно з базовими моделями, що підтверджує ефективність використання ансамблевих методів для задачі прогнозування ризику розвитку цукрового діабету. Значення Accuracy становить 0.842, що означає правильну класифікацію понад 84% випадків на тестовій вибірці. Це значно перевищує показники логістичної регресії, яка мала точність на рівні 0.767, і свідчить про здатність Random Forest краще узагальнювати інформацію з різноманітних ознак.

Метрика Precision дорівнює 0.715, що вказує на те, що серед усіх випадків, які модель класифікувала як позитивні (наявність ризику діабету), близько 71% є правильними. Це важливий показник для медичних застосувань, адже він зменшує кількість хибнопозитивних результатів, що може знизити непотрібні витрати на додаткові обстеження та психологічний стрес пацієнтів. Водночас Recall становить 0.689, що означає здатність моделі виявляти реальні позитивні випадки на рівні майже 69%. Це значно краще, ніж у логістичної регресії (0.582), і свідчить про те, що модель більш чутлива до пацієнтів із високим ризиком, що є критично важливим у контексті превентивної медицини.

Збалансованість між Precision і Recall відображає значення F1-score, яке становить 0.701. Це свідчить про гармонійне поєднання здатності моделі правильно ідентифікувати позитивні випадки та уникати помилкових спрацювань. Порівняно з логістичною регресією (F1-score = 0.602), Random Forest демонструє суттєве покращення, що підтверджує переваги

використання ансамблевих методів для задач із багатьма ознаками та потенційними нелінійними залежностями.

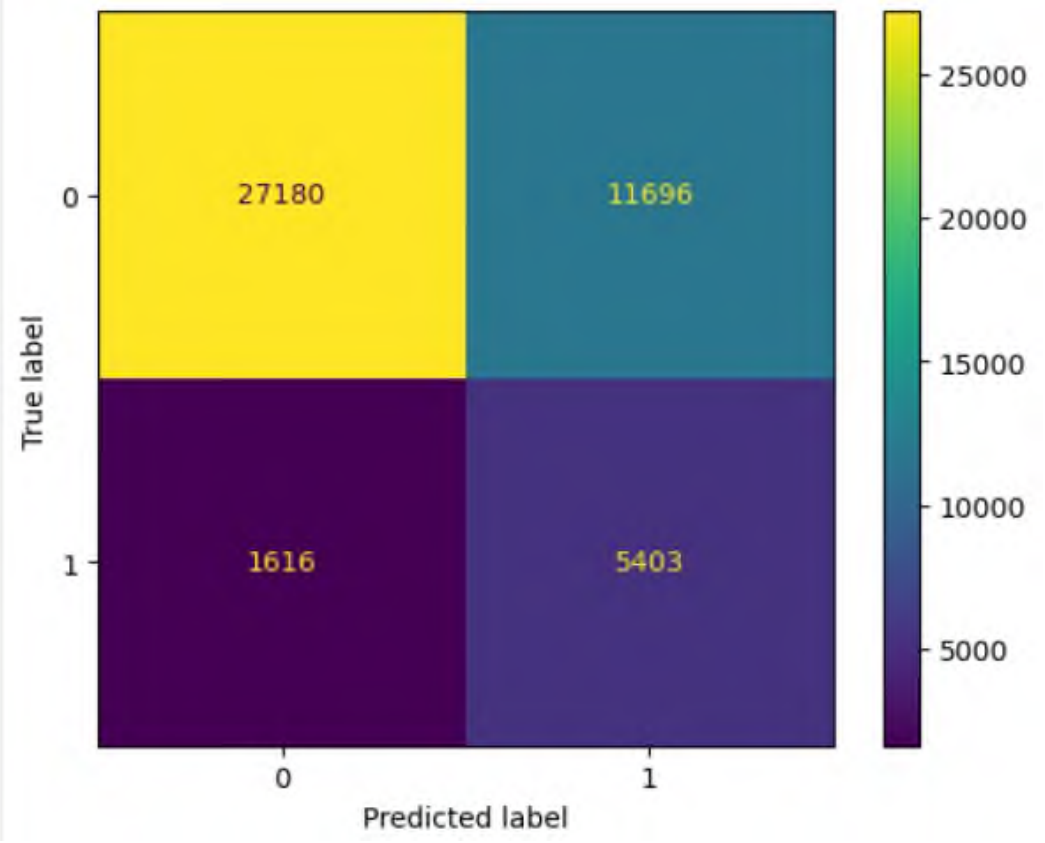

Рисунок 4.5 - Confusion матриця для Random Forest

Аналіз результатів свідчить, що Random Forest є більш ефективним рішенням для задачі прогнозування ризику цукрового діабету порівняно з логістичною регресією. Модель демонструє високу точність, чутливість та здатність до розрізнення класів, що робить її придатною для практичного використання у медичних системах. Водночас слід враховувати, що Random Forest є більш ресурсомістким алгоритмом, що потребує значних обчислювальних потужностей для навчання та прогнозування, особливо при роботі з великими наборами даних. Це може вплинути на вибір інфраструктури для розгортання моделі, зокрема на необхідність використання хмарних сервісів або спеціалізованого обладнання.

Таким чином, результати навчання Random Forest підтверджують, що оптимізація гіперпараметрів суттєво підвищує якість прогнозування, роблячи

модель потужним інструментом для превентивної медицини. Високі значення ключових метрик свідчать про готовність моделі до практичного застосування, а її гнучкість і здатність до налаштування відкривають перспективи для подальшого вдосконалення та масштабування.

4.2.3 Модель XGBoost

Після оптимізації гіперпараметрів XGBoost модель показала наступні результати

Таблиця 4.3 Результати навчання XGBoost.

Метрика	Значення
Accuracy	0.853
Precision	0.735
Recall	0.709
F1-score	0.711
ROC-AUC	0.927

Аналіз отриманих результатів показує, що модель XGBoost продемонструвала найкращу ефективність серед трьох розглянутих алгоритмів. Її точність (Accuracy) становить 0.853, що перевищує показники Random Forest (0.842) та логістичної регресії (0.767). Це свідчить про більш високу здатність XGBoost правильно класифікувати приклади. Особливо помітною є перевага за метрикою ROC-AUC — 0.927, що вказує на відмінну здатність моделі розрізняти класи навіть при зміні порогу прийняття рішення.

За показниками Precision, Recall та F1-score XGBoost також лідирує: Precision — 0.735, Recall — 0.709, F1-score — 0.711. Це означає, що модель не лише добре передбачає позитивні класи, але й зберігає баланс між точністю та повнотою. Random Forest має трохи нижчі значення (Precision — 0.715, Recall — 0.689, F1-score — 0.701), що робить його конкурентоспроможним, але менш ефективним у порівнянні з XGBoost. Логістична регресія значно поступається за всіма метриками, особливо за

Recall (0.582) та F1-score (0.602), що свідчить про її обмежену здатність виявляти позитивні випадки.

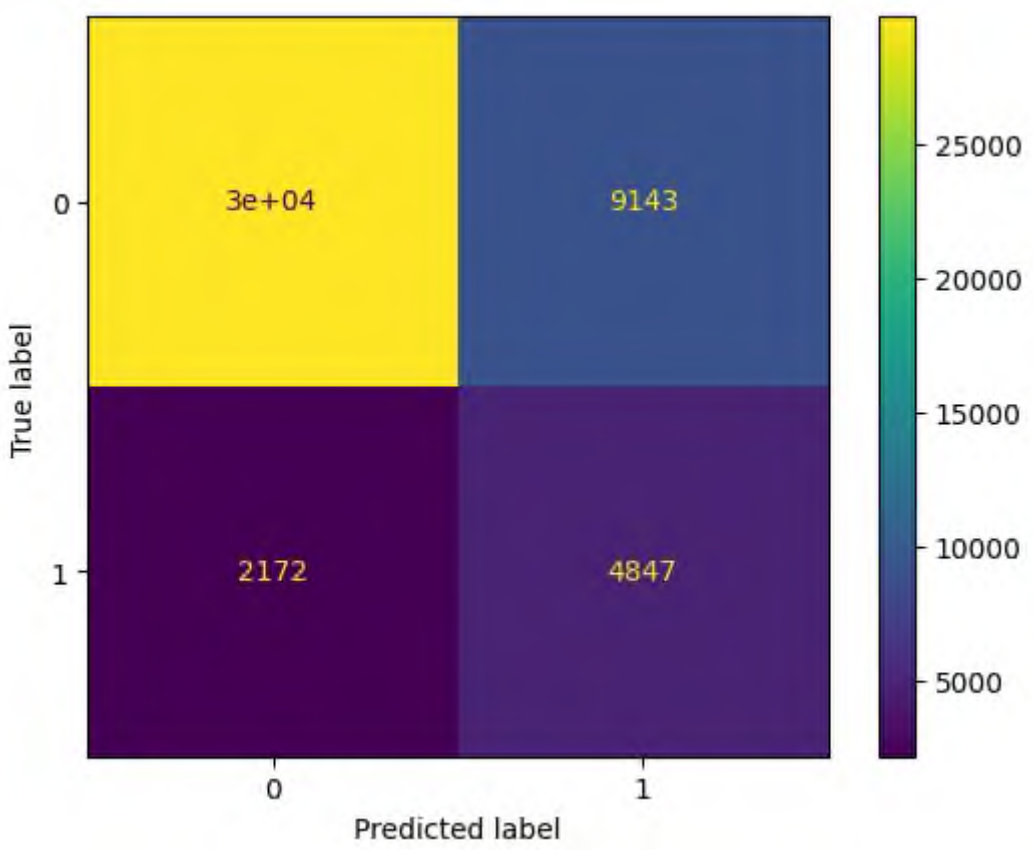

Рисунок 4.6. Confusion матриця для XGBoost

Загальний аналіз показує, що XGBoost значно перевершує логістичну регресію та Random Forest за всіма ключовими метриками, що робить його найбільш ефективним рішенням для задачі прогнозування ризику цукрового діабету. Водночас слід враховувати, що XGBoost є більш ресурсомістким алгоритмом, ніж логістична регресія, але зазвичай менш вимогливим, ніж великі ансамблі Random Forest при однаковій точності. Це робить його оптимальним вибором для систем, де важлива висока точність і чутливість, але доступні обчислювальні ресурси обмежені.

Таким чином, порівняння трьох моделей підтверджує, що використання градієнтного бустингу забезпечує найкращі результати для задачі прогнозування ризику цукрового діабету. Логістична регресія може залишатися корисною як базова модель або частина гібридного підходу,

Random Forest є хорошим компромісом між точністю та інтерпретованістю, але XGBoost демонструє найвищу ефективність і готовність до практичного застосування у системах превентивної медицини.

4.3. Висновки до розділу

Порівняння трьох моделей — XGBoost, Random Forest та логістичної регресії — показало суттєву різницю в їхній ефективності. Найкращі результати продемонструвала модель XGBoost, яка забезпечила найвищі значення точності (0.853), збалансованих метрик Precision, Recall та F1-score, а також максимальний показник ROC-AUC (0.927). Це свідчить про її здатність не лише правильно класифікувати більшість прикладів, але й ефективно розрізняти класи при зміні порогів прийняття рішень. Random Forest показав близькі, але дещо нижчі результати, що робить його конкурентоспроможним варіантом, особливо у випадках, коли важлива інтерпретованість дерев рішень. Логістична регресія значно поступається двом іншим алгоритмам за всіма метриками, що обмежує її застосування у складних задачах класифікації, хоча вона може бути корисною як базова модель завдяки простоті та прозорості.

Загалом, проведений аналіз підтверджує доцільність використання XGBoost для вирішення задачі, оскільки він забезпечує найвищу якість прогнозування та стабільність роботи. Random Forest може слугувати альтернативою при обмежених ресурсах або потребі у більш простій моделі, тоді як логістична регресія залишається базовим підходом для швидких оцінок або інтерпретованих рішень.

РОЗДІЛ 5. РОЗРОБЛЕННЯ СТАРТАП-ПРОЄКТУ

5.1. Опис ідеї проекту

Таблиця 5.1. Інформаційна карта проекту

Назва	Опис
Вид проекту	Інтелектуальна система
Назва проекту	«Інтелектуальна система прогнозування появи цукрового діабету»
Назва ВНЗ в якому розробляється проект	Національний лісотехнічний університет України, кафедра комп'ютерних наук
Прізвище, ім'я, побатькові	Василишин Н.Т.
Цілі та задачі проекту	1. На основі бази знань, що складається із класифікованих даних появи цукрового діабету, створити нейромережеву модель. 2. Розробити механізм емуляції нейронної мережі та прогнозування появи цукрового діабету. 3. Розроблена система повинна мати інтуїтивно зрозумілий інтерфейс.
Короткий зміст проекту	Проект покликаний створити інтелектуальну систему прогнозування появи цукрового діабету. Це дозволить автоматизувати процес ідентифікації ризику появи цукрового діабету.
Терміни виконання проекту	6 місяців
Бюджет проекту	183 000 грн.

5.2. Ідея та концепція стартапу

Ідея стартапу ґрунтується на створенні інтелектуальної системи, здатної прогнозувати ризик виникнення цукрового діабету на ранніх етапах, коли клінічні прояви ще відсутні, але існують приховані фактори, що можуть призвести до розвитку захворювання. Цукровий діабет є однією з найпоширеніших хронічних хвороб сучасності, яка має тенденцію до стрімкого зростання у всьому світі. За даними міжнародних організацій, кількість людей із діабетом щороку збільшується, а економічні та соціальні наслідки цього процесу стають дедалі відчутнішими. Проблема полягає не лише у високій вартості лікування, але й у значному зниженні якості життя пацієнтів, що стикаються з ускладненнями, такими як серцево-судинні захворювання, нефропатія, ретинопатія та інші. Традиційні методи діагностики здебільшого орієнтовані на виявлення хвороби після її розвитку, тоді як сучасні підходи повинні бути спрямовані на превентивні заходи, що дозволяють запобігти виникненню патології.

Концепція стартапу передбачає використання алгоритмів машинного навчання для аналізу великого масиву даних, що включає медичні показники, інформацію про спосіб життя, генетичні особливості та поведінкові фактори. На основі цих даних система формує прогноз ризику розвитку діабету та надає персоналізовані рекомендації щодо профілактики. Такий підхід дозволяє не лише підвищити точність прогнозування, але й зробити його доступним для широкого кола користувачів через інтеграцію з мобільними додатками та веб-платформами. Унікальність ідеї полягає в поєднанні медичних знань із сучасними технологіями штучного інтелекту, що відкриває нові можливості для охорони здоров’я.

Актуальність проєкту обумовлена глобальними тенденціями зростання кількості хворих на діабет. За прогнозами Всесвітньої організації охорони здоров’я, до 2030 року кількість пацієнтів із цукровим діабетом може перевищити 500 мільйонів осіб. Це створює значне навантаження на системи охорони здоров’я та економіку країн. Водночас більшість випадків діабету другого типу можна попередити за умови своєчасного виявлення ризиків і

корекції способу життя. Саме на цьому базується концепція нашого стартапу: надати користувачам інструмент, який дозволить оцінити ймовірність розвитку хвороби та отримати рекомендації щодо її запобігання.

Реалізація проєкту передбачає створення зручного та інтуїтивно зрозумілого інтерфейсу, який дозволить користувачам легко взаємодіяти із системою. Особлива увага приділятиметься безпеці даних, адже йдеться про конфіденційну медичну інформацію. Використовуватимуться сучасні протоколи шифрування та захисту персональних даних відповідно до міжнародних стандартів. Це забезпечить довіру користувачів і відповідність нормативним вимогам.

Таким чином, ідея та концепція стартапу базуються на інтеграції медичних знань, технологій штучного інтелекту та принципів превентивної медицини. Проєкт має потенціал для значного впливу на охорону здоров'я, економіку та якість життя людей. Його реалізація сприятиме зниженню поширеності цукрового діабету, оптимізації витрат на лікування та формуванню нової культури відповідального ставлення до власного здоров'я.

5.3. Технологічна реалізація

Технологічна реалізація стартапу «Інтелектуальна система прогнозування появи цукрового діабету» ґрунтується на використанні сучасних методів обробки даних, алгоритмів машинного навчання та хмарних технологій для забезпечення масштабованості та доступності рішення. Основна мета технологічної частини полягає у створенні надійної, безпечної та високопродуктивної платформи, яка здатна аналізувати великі обсяги медичних і поведінкових даних, формувати точні прогнози та надавати користувачам персоналізовані рекомендації в режимі реального часу.

Архітектура системи передбачає багаторівневу структуру, що включає модулі збору даних, їх попередньої обробки, аналітичного ядра та інтерфейсу взаємодії з користувачем. Збір даних здійснюється з різних джерел: електронних медичних карток, результатів лабораторних досліджень, даних

із носимих пристроїв, а також інформації, яку користувач вводить самостійно через мобільний додаток або веб-платформу. Для забезпечення сумісності з медичними системами використовуються стандарти обміну даними HL7 та FHIR, що дозволяє інтегрувати рішення у вже існуючу інфраструктуру охорони здоров'я.

Попередня обробка даних є критично важливим етапом, оскільки медичні дані часто містять пропуски, помилки або мають різні формати. Для цього застосовуються методи нормалізації, заповнення пропусків, а також алгоритми виявлення аномалій. Особлива увага приділяється забезпеченню конфіденційності та захисту персональних даних. Використовуються сучасні протоколи шифрування, а доступ до інформації регламентується відповідно до міжнародних стандартів, таких як GDPR та HIPAA. Це дозволяє гарантувати безпечне зберігання та передачу даних між компонентами системи.

Аналітичне ядро платформи базується на алгоритмах машинного навчання, які дозволяють формувати прогнози ризику розвитку цукрового діабету. Для цього використовуються моделі, що поєднують класичні статистичні методи та сучасні підходи, такі як логістична регресія, дерева рішень, ансамблеві методи (Random Forest, Gradient Boosting) та нейронні мережі. Гібридний підхід забезпечує високу точність прогнозів, оскільки враховує як медичні показники, так і поведінкові фактори, включаючи рівень фізичної активності, харчові звички та стресові навантаження. Навчання моделей здійснюється на великих масивах даних, що дозволяє враховувати різноманітні комбінації факторів ризику та підвищувати адаптивність системи.

Для реалізації аналітичних алгоритмів обрано мову програмування Python, яка є стандартом у сфері Data Science та має широкий набір бібліотек для роботи з даними та побудови моделей машинного навчання. Використовуються такі інструменти, як TensorFlow та PyTorch для нейронних мереж, Scikit-learn для класичних алгоритмів, а також Pandas і NumPy для обробки даних. Це забезпечує гнучкість у розробці та можливість

швидкої адаптації до нових вимог. Крім того, застосовується контейнеризація за допомогою Docker, що дозволяє легко розгортати систему на різних платформах та забезпечує її масштабованість.

Інтерфейс взаємодії з користувачем реалізується у вигляді мобільного додатку та веб-платформи. Мобільний додаток дозволяє користувачам вводити дані, отримувати прогнози та рекомендації, а також синхронізувати інформацію з носимими пристроями. Веб-платформа орієнтована на медичних працівників і страхові компанії, надаючи їм доступ до аналітичних звітів та інструментів для управління ризиками. Особлива увага приділяється зручності інтерфейсу та його адаптивності для різних пристроїв, що забезпечує позитивний досвід користувача.

Хмарна інфраструктура є ключовим елементом технологічної реалізації, оскільки вона дозволяє обробляти великі обсяги даних та забезпечує доступність сервісу для користувачів у будь-якій точці світу. Використовуються сервіси провідних постачальників, таких як AWS або Microsoft Azure, що гарантує високу продуктивність, надійність та можливість масштабування. Хмарні технології також забезпечують резервне копіювання даних та їх відновлення у разі збою, що є критично важливим для медичних систем.

Важливим аспектом є тестування та валідація моделей прогнозування. Для цього застосовуються методи крос-валідації, а також порівняння результатів із реальними клінічними даними. Це дозволяє оцінити точність та надійність системи перед її впровадженням у практику. Крім того, передбачено механізм постійного оновлення моделей на основі нових даних, що забезпечує їх актуальність та підвищує ефективність прогнозів.

Таким чином, технологічна реалізація стартапу поєднує сучасні методи обробки даних, алгоритми машинного навчання, хмарні технології та зручні інтерфейси для користувачів. Вона забезпечує високу точність прогнозів, безпеку даних та доступність сервісу, що робить проєкт конкурентоспроможним і перспективним для впровадження у сфері охорони здоров’я.

5.4. Фінансовий план

Фінансовий план стартапу є ключовим елементом стратегії реалізації проєкту, оскільки він визначає обсяг необхідних ресурсів, джерела фінансування, структуру витрат та прогнозовані доходи. Розробка фінансової моделі ґрунтується на принципах реалістичності, гнучкості та орієнтації на довгострокову окупність. Враховуючи специфіку медичних технологій, фінансовий план передбачає значні початкові інвестиції, пов'язані з розробкою програмного забезпечення, тестуванням, сертифікацією та маркетинговими заходами. Однак ці витрати компенсуються високим потенціалом комерціалізації продукту та можливістю масштабування на міжнародному ринку.

Першим етапом фінансового планування є визначення стартових витрат, які охоплюють розробку аналітичного ядра системи, створення мобільного додатку та веб-платформи, а також забезпечення хмарної інфраструктури для обробки даних. Значну частину бюджету займають витрати на оплату праці команди розробників, спеціалістів з машинного навчання, UX/UI дизайнерів та експертів у сфері охорони здоров'я. Крім того, необхідно врахувати витрати на ліцензування програмних компонентів, придбання серверних потужностей та забезпечення інформаційної безпеки відповідно до міжнародних стандартів. Важливим аспектом є фінансування процесу тестування та валідації моделей прогнозування, що потребує залучення клінічних даних і співпраці з медичними установами.

Окрему статтю витрат становить маркетинг і просування продукту. Для виходу на ринок необхідно сформувати ефективну комунікаційну стратегію, яка включає розробку бренду, створення рекламних матеріалів, проведення інформаційних кампаній у соціальних мережах та організацію презентацій для потенційних партнерів. Враховуючи специфіку продукту, особливу увагу слід приділити співпраці з медичними закладами, страховими компаніями та організаціями, що займаються профілактикою хронічних захворювань. Це

потребує додаткових витрат на участь у галузевих конференціях, виставках та проведення демонстраційних проєктів.

Джерела фінансування стартапу можуть бути різноманітними. На початковому етапі доцільно залучити грантові програми, спрямовані на підтримку інновацій у сфері охорони здоров'я, а також венчурні інвестиції від фондів, що спеціалізуються на медичних технологіях. Додатковим джерелом фінансування може стати краудфандинг, який дозволяє не лише отримати кошти, але й сформувати спільноту користувачів, зацікавлених у продукті. У перспективі можливе залучення стратегічних партнерів серед великих медичних компаній або страхових організацій, які зацікавлені у впровадженні технологій прогнозування ризиків для оптимізації своїх витрат.

Окупність проєкту прогнозується протягом 18–24 місяців після запуску комерційної версії продукту. Це обумовлено високим попитом на рішення, що дозволяють знижувати ризики розвитку хронічних захворювань, а також можливістю масштабування продукту на міжнародному ринку. Важливим фактором є гнучкість фінансової моделі, яка дозволяє адаптувати стратегію монетизації залежно від ринкових умов. Наприклад, у разі високої конкуренції на ринку мобільних додатків можна посилити акцент на корпоративному сегменті, де бар'єри входу є вищими, але й потенційні доходи значно більші.

Ризики, пов'язані з фінансовим планом, включають можливі затримки у розробці продукту, необхідність додаткових витрат на сертифікацію та регуляторні вимоги, а також коливання попиту на ринку. Для мінімізації цих ризиків передбачено створення резервного фонду, який дозволить покрити непередбачені витрати, а також розробку сценаріїв адаптації бізнес-моделі. Крім того, важливим завданням є постійний моніторинг фінансових показників і корекція стратегії залежно від результатів.

Таким чином, фінансовий план стартапу є комплексною системою, що охоплює всі аспекти реалізації проєкту – від стартових інвестицій до прогнозування доходів і управління ризиками. Його реалізація забезпечує стабільність розвитку, можливість масштабування та досягнення

стратегічних цілей, спрямованих на комерціалізацію інноваційного продукту у сфері охорони здоров’я.

5.5. Висновки до розділу

Розроблення стартапу «Інтелектуальна система прогнозування появи цукрового діабету» демонструє комплексний підхід до вирішення актуальної медичної проблеми шляхом інтеграції сучасних технологій штучного інтелекту, хмарних сервісів та принципів превентивної медицини. У межах розділу було сформовано концепцію проєкту, визначено цільову аудиторію, проведено аналіз ринку та конкурентів, розроблено бізнес-модель, описано технологічну реалізацію, фінансовий план і ризики. Запропоноване рішення має значний соціальний та економічний потенціал, оскільки спрямоване на зниження поширеності хронічних захворювань, оптимізацію витрат на лікування та підвищення якості життя людей. Реалізація проєкту забезпечить створення інноваційного продукту, здатного зайняти конкурентну позицію на ринку цифрової медицини та стати основою для подальшого розвитку екосистеми превентивних технологій.

ВИСНОВКИ

У межах цієї роботи було проведено комплексне дослідження щодо розробки та оцінювання інтелектуальної системи прогнозування появи цукрового діабету на основі сучасних методів машинного навчання.

Актуальність проблеми цукрового діабету підтверджується світовими статистичними даними, які свідчать про стале зростання кількості хворих. У цьому контексті раннє виявлення факторів ризику із застосуванням цифрових технологій має надзвичайно високу практичну цінність.

Було проаналізовано існуючі підходи до виявлення та прогнозування діабету, зокрема клініко-лабораторні методи, експертні системи та моделі машинного навчання. Установлено, що моделі штучного інтелекту здатні забезпечити високу точність прогнозування, особливо при роботі з великими обсягами структурованих даних.

В рамках інформаційного забезпечення дослідження було обрано датасет BRFSS 2015, який містить ключові соціально-демографічні, поведінкові та медичні індикатори, релевантні для оцінки ризику розвитку діабету. Проведено попередню обробку даних, включно з очищенням, нормалізацією та балансуванням вибірки, що є критичним етапом для побудови коректних моделей машинного навчання.

Побудовано та навчено дві моделі: логістичну регресію та Random Forest. Здійснено глибокий аналіз результатів, зокрема порівняння їх точності, повноти, F1-міри та ROC-AUC. Встановлено, що модель Random Forest продемонструвала кращу загальну продуктивність, хоча логістична регресія має вищу інтерпретованість.

Розроблена інтелектуальна система може бути використана як допоміжний інструмент для лікарів, епідеміологів, страхових компаній чи органів охорони здоров’я для скринінгу ризику захворювання, автоматизованої обробки великих масивів даних і виявлення найбільш уразливих груп населення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Василишин Н, Шиманський В. Інтелектуальна система прогнозування появи цукрового діабету. КМІТ – 2025, Львів, 2025.

2. Amjad M., Ali Z., Rafiq A., Akhtar N., - I.-U.-R., Abbas A. Empirical Performance Analysis of Decision Tree and Support Vector Machine based Classifiers on Biological Databases. International Journal of Advanced Computer Science and Applications. 2019. Вип. 10, № 9.

3. Nuankaew P., Chaising S., Temdee P. Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction. IEEE Access. 2021. Вип. 9. С. 137015–137028.

4. Sabitha E., Durgadevi M. Improving the Diabetes Diagnosis Prediction Rate Using Data Preprocessing, Data Augmentation and Recursive Feature Elimination Method. International Journal of Advanced Computer Science and Applications. 2022. Вип. 13, № 9.

5. Tan Y., Chen H., Zhang J., Tang R., Liu P. Early Risk Prediction of Diabetes Based on GA-Stacking. Applied Sciences. 2022. Вип. 12, № 2. С. 632.

6. Alghamdi T. Prediction of Diabetes Complications Using Computational Intelligence Techniques. Applied Sciences. 2023. Вип. 13, № 5. С. 3030.

7. Fazakis N., Kocsis O., Dritsas E., Alexiou S., Fakotakis N., Moustakas K. Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction. IEEE Access. 2021. Вип. 9. С. 103737–103757.

8. Ferdousi R., Hossain M. A., El Saddik A. Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS. IEEE Access. 2021. Вип. 9. С. 96823–96837.

9. Syed A. H., Khan T. Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study. IEEE Access. 2020. Вип. 8. С. 199539–199561.

10. ElSeddawy A. I., Karim F. K., Hussein A. M., Khafaga D. S. Predictive Analysis of Diabetes-Risk with Class Imbalance. Computational Intelligence and Neuroscience. 2022. Вип. 2022. С. 1–16.

11. Bhat S. S., Selvam V., Ansari G. A., Ansari M. D., Rahman M. H. Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora. Computational Intelligence and Neuroscience. 2022. Вип. 2022. С. 1–12.

12. Saxena R., Sharma S. K., Gupta M., Sampada G. C. A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. Computational Intelligence and Neuroscience. 2022. Вип. 2022.

ДОДАТОК А

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv('diabetes_binary_health_indicators_BRFSS2015.csv')
df = data.copy()

df.drop_duplicates(inplace = True)
target = ['Diabetes_binary']
features_binary = ['HighBP', 'HighChol', 'CholCheck', 'Smoker', 'Stroke', 'HeartDiseaseorAttack',
                   'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost',
'DiffWalk', 'Sex']
features_ordinal = ['GenHlth', 'Age', 'Education', 'Income']
features_numerical = ['BMI', 'MentHlth', 'PhysHlth']

# Set up subplots for plotting
fig, axes = plt.subplots(nrows=5, ncols=3, figsize=(15, 15))
fig.subplots_adjust(hspace=0.5)

# Loop through binary features and create plots
for i, feature in enumerate(features_binary):
    row, col = i // 3, i % 3
    sns.countplot(x=feature, hue='Diabetes_binary', data=df, ax=axes[row, col], palette='pastel')
    axes[row, col].set_title(f'Distribution of {feature}')

plt.tight_layout()
plt.show()

fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(12,7))
fig.subplots_adjust(hspace=0.5)

# Loop through binary features and create plots
for i, feature in enumerate(features_ordinal):
    row, col = i // 2, i % 2
    sns.countplot(x=feature, hue='Diabetes_binary', data=df, ax=axes[row, col], palette='muted')
    axes[row, col].set_title(f'Distribution of {feature}')

plt.tight_layout()
plt.show()

plt.figure(figsize=(15, 8))

for i, feature in enumerate(features_numerical, 1):
    plt.subplot(2, 3, i)
    sns.histplot(df, x=feature, kde=True, bins=20, hue='Diabetes_binary', multiple='stack', palette='deep')
    plt.title(f'Distribution of {feature}')

    plt.subplot(2, 3, i + 3)
    sns.boxplot(x=df[feature], color='plum')
    plt.title(f'Boxplot of {feature}')

plt.tight_layout()
plt.show()

plt.figure(figsize=(15, 8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
X = df.drop(columns = 'Diabetes_binary')
```

```
y = df['Diabetes_binary']

print(X.shape, y.shape)

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=88, stratify = y)

print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

from sklearn.preprocessing import StandardScaler, PolynomialFeatures
from sklearn.pipeline import Pipeline

from sklearn.model_selection import cross_validate, cross_val_score, cross_val_predict
from sklearn.model_selection import StratifiedKFold
from sklearn.experimental import enable_halving_search_cv
from sklearn.model_selection import GridSearchCV, HalvingGridSearchCV
from sklearn.metrics import classification_report, confusion_matrix, ConfusionMatrixDisplay
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
balanced_accuracy_score, roc_auc_score

import warnings
from sklearn.exceptions import ConvergenceWarning

# Ignore ConvergenceWarnings
warnings.filterwarnings("ignore", category=ConvergenceWarning)

def measure_error(y_true, y_pred, label):
    return pd.Series({
        'balanced_accuracy': balanced_accuracy_score(y_true, y_pred),
        'precision': precision_score(y_true, y_pred, average='weighted'),
        'recall': recall_score(y_true, y_pred, average='weighted'),
        'f1': f1_score(y_true, y_pred, average='weighted'),
        'auc_roc': roc_auc_score(y_true, y_pred, average='weighted')
    }, name=label)


# Function to evaluate the model - fits the model, then computes training vs testing set metrics
def evaluate_model(model, X_train, y_train, X_test, y_test):
    model.fit(X_train, y_train)
    y_train_pred = model.predict(X_train)
    y_test_pred = model.predict(X_test)

    # Computes the metrics for the training & testing set, and presents them in a dataframe
    train_test_error = pd.concat([measure_error(y_train, y_train_pred, 'Train'),
                                  measure_error(y_test, y_test_pred, 'Test')],
                                  axis=1)
    print(train_test_error, "\n")

    # Display classification report for test set
    print("Testing Set Classification Report:\n", classification_report(y_test, y_test_pred))

    # Display confusion matrix for the Test set
    cm = confusion_matrix(y_test, y_test_pred)
    disp = ConfusionMatrixDisplay(cm)

    return disp.plot()

from sklearn.linear_model import LogisticRegression

# Initialise standard scaler
scaler = StandardScaler()

# Create a logistic regression model without regularisation
```

```
logr = LogisticRegression()

# Create a pipeline with a standard scaler and the logistic regression model
pipe = Pipeline([
   #(nickname, step)
   ('scaler', scaler),
   ('logr', logr)
])

# Evaluate the model
evaluate_model(pipe, X_train, y_train, X_test, y_test)

rfc = RandomForestClassifier(
   n_estimators = 100,  # default
   max_features='sqrt',
   # max_depth= ?, # to optimise
   # max_leaf_nodes = ?,  # to optimise
   # min_samples_split = ? # to optimise
   class_weight='balanced', # ensure balanced weights to account for class imbalance (balanced, balanced_subsample - very similar results)
   random_state=88)

### Parameters to search over ###
param_grid = {
   #'step__param' : [list of param values]
   'max_depth': [15, 20, 25, 30],
   'max_leaf_nodes': [50, 75, 100, 150],
   'min_samples_split': [50, 100, 200, 400]
}

### Validator ###
# Use 3 k-fold, considering the size of the dataset & computational power
cv = StratifiedKFold(n_splits=3, shuffle=True, random_state=88)

## GridSearch ###
grid = GridSearchCV(rfc,       # model to use
                    param_grid, # parameters to search over
                    scoring = ['precision', 'recall', 'f1', 'balanced_accuracy', 'roc_auc'], # metrics to compute
                    refit = 'balanced_accuracy',  # which metric to use to decide the best model
                    cv = cv,
                    n_jobs = -1
                   )

grid.fit(X_train, y_train)
```