КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТІ ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій
Кафедра технологій управління

Спеціальність 122 «Комп'ютерні науки»
Освітня програма «Інформаційна аналітика та впливи»

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:

"Технологія визначення діабету методами машинного навчання"

Студента 2-го курсу групи ІАВ-21

Федірко Юлії Ігорівни
(прізвище, ім'я, по батькові)

(підпис студента)

Науковий керівник:

д.т.н., професор.
(науковий ступінь, вчене звання)

Хлевна Юлія Леонідівна
(прізвище, ім'я, по батькові)

(дата) (підпис)

Попередній захист:

(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри
технологій управління

(підпис) (прізвище, ініціали) (дата)

Київ – 2023

ЗАВДАННЯ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій

Кафедра технологій управління
Освітньо-кваліфікаційний рівень Магістр
Спеціальність 122 - Комп'ютерні науки
Освітня програма Інформаційна аналітика та впливи

ЗАТВЕРДЖУЮ
Завідувач кафедри
професор Морозов В.В.

_____
«____» _____20___року

З А В Д А Н Н Я
НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студент Федірко Юлія Ігорівна

Група ІАВ-21

1. Тема кваліфікаційної роботи

Технологія визначення діабету методами машинного навчання

Затверджена наказом по від «08»12 2022 р. № 5 .

2. Строк подання студентом готової роботи – "18"05 2023 р.

3. Цільова установка та вихідні дані до роботи

Цільова установа - ТОВ «Інфопульс», використано дані зі сховища машинного навчання UCI, а також набір даних іракською лабораторією медичної міської лікарні.

4. Зміст роботи

У роботі досліджуються існуючі підходи до використання методів машинного навчання до проблеми визначення діабету другого типу. Розробляється нова технологія їх використання, а також проводиться обґрунтування доцільності та необхідності впровадження запропонованої технології. Наводяться рекомендації щодо практичної імплементації технології.

5. Перелік графічного матеріалу (слайдів)
Зведена характеристика про захворювання ДДТ, Динаміка запитів щодо симптомів діабету, Дерево проблем та дерево рішень, Порівняльна характеристика конкурентів, Баланс між ризиками та рішеннями питання впровадження машинного навчання до медичної сфери, Дескриптивний аналіз набору даних, Кореляційний аналіз ознак набору даних з результуючою ознакою клас, Приклад дерева для ознак, Графічне порівняння метрик методів, Результати створеної моделі, Архітектура рішення, SWOT аналіз проекту

6. Календарний план виконання роботи:

№ п/п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1.	Вибір теми дипломної роботи	3	01.10.22	01.10.22
2.	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників	2	27.12.22	27.12.22
3.	Формування переліку нормативних матеріалів, літератури з проблематики дипломної роботи	10	08.01.23	07.01.23
4.	Складання розгорнутого плану кваліфікаційної роботи	5	18.01.23	18.01.23
5.	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін.	5	19.01.23- 20.01.23	20.11.23
6.	Підготовка розділу 1 «Аналіз теоретико-методологічних основ використання інформаційної аналітики в медицині»	10	12.02.23	13.02.23
7.	Підготовка розділу 2 «ММН для визначення діабету»	14	08.03.23	08.03.23
8.	Підготовка розділу 3 «Побудова моделі прогнозування діабету за допомогою методу випадкового лісу»	14	01.04.23	01.04.23
9.	Підготовка розділу 4 «Застосування моделі прогнозування ДДТ»	13	20.04.23	20.04.23
10.	Оформлення кваліфікаційної роботи. Підготовка висновків і пропозицій	15	03.05.23	03.05.23
11.	Передача кваліфікаційної роботи науковому керівникові	2	04.05.23	04.05.23
12.	Передача кваліфікаційної роботи рецензенту для рецензування	2	11.05.23	11.05.23
13.	Попередній захист кваліфікаційної роботи	5	17.05.23	17.05.23

Дата видачі завдання «08» 12 2022 р.

Керівник роботи д.т.н., професор. Хлевна Юлія Леонідівна
(посада, прізвище, ім’я, по батькові)

_____
(підпис)

Завдання прийняв до виконання студент групи ІАВ-21

Федірко Юлія Ігорівна
(прізвище, ім’я, по батькові)

_____
(підпис)

ЗМІСТ

АНОТАЦІЯ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА Ш ЕВЧЕНКА
Факультет інформаційних технологій
Кафедра технологій управління
Спеціальність 122 - Комп'ютерні науки,
освітня програма "Інформаційна аналітика та впливи"

Дипломна робота магістра Федірко Юлії Ігорівни.

Тема роботи – «Технологія визначення діабету методами машинного навчання».

Мета дипломної роботи магістра – розробити технологію визначення діабету другого типу методом машинного навчання.

Об'єкт дослідження – процеси визначення діабету другого типу.

Предмет дослідження – методики застосування машинного навчання у задачах визначення діабету другого типу.

Наукова новизна роботи – розроблено технологію визначення діабету другого типу на основі методу випадкового лісу, яка відрізняється від існуючих тим, що дозволяє визначати діабет на основі як категоріальних, так і числових даних, а також вищою точністю. Запропоновано проект застосування моделі прогнозування діабету другого типу у вигляді Power App застосунку з інтеграцією Microsoft Data Driven застосунками, а також запропоновано шаблони аналітичної звітності, структури бази даних. Створено архітектуру рішення.

У роботі досліджуються існуючі підходи до використання методів машинного навчання до проблеми визначення діабету другого типу. Розробляється нова технологія їх використання, а також проводиться обґрунтування доцільності та необхідності впровадження запропонованої технології. Наводяться рекомендації щодо практичної імплементації технології.

Дипломна робота складається зі вступу, основної частини, яка включає чотири розділи, висновків та списку використаних джерел. Всього налічує 99 сторінок та перелік посилань з 74 джерел на 8 сторінках.

Ключові слова: технологія, інтелектуальний аналіз даних, метод випадкового лісу, методи машинного навчання, діабет другого типу.

ПЕРЕЛІК ВИКОРИСТАНИХ СКОРОЧЕНЬ

ММН – методи машинного навчання

ДДТ – діабет другого типу

ВСТУП

Актуальність роботи.

Діабет — це хронічний метаболічний розлад, що характеризується високим рівнем цукру в крові, що визначається недостатнім виробництвом інсуліну, гормону, який виробляє підшлункова залоза, який регулює поглинання та метаболізм глюкози, основного джерела енергії для клітин організму.

Діабет другого типу – є гетерогенним захворюванням, що часто непросто класифікувати у пацієнтів. Люди з встановленим захворюванням діабету знаходяться в групі ризику розвитку хронічних ускладнень, таких як хвороба нирок, серцева хвороба, інсульт, пошкодження нервів і втрата зору. Рання діагностика та лікування, включаючи зміну способу життя та прийом ліків, можуть запобігти або відстрочити появу ускладнень, зокрема при діабеті другого типу.

Цукровий діабет входить до десятки основних причин смерті в усьому світі. Відповідно до 10-го Атласу діабету IDF, у 2021 році глобальна поширеність діабету серед людей віком 20–79 років становила 10,5% (536,6 мільйонів осіб), а за прогнозами, у 2045 році вона зросте до 12,2% (783,2 мільйона). У наш час лікування діабету все ще є проблемою, майже кожна друга доросла людина, яка страждає на цю патологію, не знає про свій статус.

Методи машинного навчання набули своєї популярності через те, що показали гарні результати у визначенні різноманітних захворювань, в тому числі і діабету другого типу. Адже методи машинного навчання можуть ефективно обробляти великі обсяги медичних даних, включаючи клінічні записи пацієнтів, результати лабораторних тестів та інші фактори здоров'я. Це дозволяє виявити складні взаємозв'язки та кореляції між різними показниками та ризиком розвитку діабету.

Моделі машинного навчання можуть досягати високої точності при класифікації та прогнозуванні. Вони можуть враховувати багато факторів одночасно та здатні виявляти навіть незначні зміни у показниках здоров'я, які

можуть бути пов'язані з розвитком діабету.

Машинне навчання може виявити комплексні патерни та взаємозв'язки, які можуть бути складні для сприйняття людиною. Моделі машинного навчання можуть виявити навіть слабкі асоціації та ризикові фактори, що допомагають точніше визначити наявність діабету.

Питанням прогнозування діабету займалися як вітчизняні так й іноземні дослідники, зокрема: Ханг Лай, Хуасюн Хуан, Карім Кешавджі, Азіз Гергачі та Сінь Гао, Харлін Каур, Вініта Кумарі, Неха Прерна Тіггаа, Шруті Гарг, О.С. Кротова, М.Р. Басараб та інші.

Об’єктом дослідження є процеси визначення діабету другого типу.

Предметом дослідження методики застосування машинного навчання у задачах визначення діабету другого типу.

Мета роботи. Метою цього дослідження є розроблення технології визначення діабету другого типу методом машинного навчання.

Постановка мети зумовила необхідність виокремлення і вирішення таких завдань:

1) Проаналізувати теоретико-методологічні основи використання інформаційної аналітики в медицині, зокрема проаналізувати процеси визначення діабету другого типу, використання методів машинного навчання на реальних підприємствах в сфері визначення діабету другого типу та визначення необхідності застосування методів машинного навчання в даній галузі.
2) Дослідити алгоритми машинного навчання загалом та в сфері виявлення діабету другого типу, зробити математичний опис методів, визначити найбільш підходящий метод машинного навчання для досягнення мети роботи.
3) Побудувати технологію прогнозування діабету другого типу за допомогою обраного метода машинного навчання, а саме формалізувати базу знань захворювання діабету другого типу, обрати

технічні засоби реалізації, та побудувати технологію машинного навчання.

4) Запропонувати практичну реалізацію створеної технології визначення діабету, описати проект та його розгортання а також структурні елементи запропонованого проекту.

Інформаційними джерелами є наукові роботи вітчизняних та іноземних авторів, довідкові сайти, документація відповідних бібліотек.

Методи дослідження. У роботі для розкриття поставлених завдань використано загальнонаукові та загальні методи дослідження. В межах першого та другого розділу основними методами є методи теоретичного дослідження: абстрагування та методи дедукції та індукції, методи аналізу і синтезу використано для формування порівняльної таблиці існуючих рішень проблеми визначення ДДТ.

В межах третього розділу використано такі методи які включають збір даних, вибір моделі, розробку моделі та оцінку розробленої моделі. Дані будуть зібрані з відповідних джерел, таких як медичні записи пацієнтів, а розробка моделі включатиме вибір і перетворення найважливіших характеристик у даних для створення набору значущих змінних для моделі. Вибір моделі включатиме вибір відповідного алгоритму машинного навчання для завдання, а систему буде оцінено за допомогою відповідних показників, таких як accuracy, precision, recall, and F1-score. Вибраним методом для цього дослідження буде алгоритм випадкового лісу через його доведену здатність добре виконувати завдання класифікації зі складними наборами даних.

В межах четвертого рекомендаційного розділу використано методи моделювання, для відображення архітектури запропонованого рішення, також для надання рекомендацій використано методи сходження від абстрактного до конкретного.

Також в процесі роботи використовуються спеціальні методи наукового дослідження, такі як SWOT-аналіз, PEST-аналіз, методом Power/Interest Grid тощо.

Наукова новизна одержаних результатів. Розроблено технологію визначення діабету другого типу на основі методу випадкового лісу, яка відрізняється від існуючих тим, що дозволяє визначати діабет на основі як категоріальних, так і числових даних, а також вищою точністю аніж у останніх джерелах. Удосконалено алгоритм виявлення діабету з використанням виокремлення основних незалежних змінних за допомогою кореляції та методу головних компонент.

Практичне значення одержаних результатів: Запропоновано проект застосування моделі прогнозування діабету другого типу у вигляді Power App застосунку з інтеграцією Microsoft Data Driven застосунками, а також запропоновано шаблони аналітичної звітності, структури бази даних. Створено архітектуру рішення. Використання комбінації методів машинного навчання та зберігання даних у формі зірки, хмарних технологій та low-code застосунку.

Апробація результатів. Результати дослідження були оприлюдненні у матеріалах конференції «Інформаційні технології та впровадження» 2021 року у роботі Федірко Ю., Єгорченков О. Алгоритм випадкового лісу в прогнозуванні цукрового діабету (тип 2) та 2022 року у роботі Федірко Ю., Єгорченков О. Модель глибокого навчання в прогнозуванні діабету (Тип 2) [25,24].

РОЗДІЛ 1. АНАЛІЗ ТЕОРЕТИКО-МЕТОДОЛОГІЧНИХ ОСНОВ ВИКОРИСТАННЯ ІНФОРМАЦІЙНОЇ АНАЛІТИКИ В МЕДИЦИНІ

1.1. Аналіз об’єкту дослідження

ДДТ – це хронічне захворювання, яке виникає, коли підшлункова залоза не виробляє достатньо інсуліну, або коли організм не може ефективно використовувати інсулін, який він виробляє. Інсулін - це гормон, який регулює рівень цукру в крові. Гіперглікемія, або підвищений рівень цукру в крові, є поширеним наслідком неконтрольованого діабету і з часом призводить до серйозних пошкоджень багатьох систем організму, особливо нервів і кровоносних судин.

Цукровий діабет 2 типу виникає внаслідок неефективного використання інсуліну організмом. Більше 95% людей з цукровим діабетом мають цукровий діабет 2 типу. Цей тип цукрового діабету багато в чому є результатом надмірної маси тіла і гіподинамії. Донедавна цей тип цукрового діабету спостерігався тільки у дорослих, але зараз він все частіше зустрічається і у дітей.

Симптоми ДДТ не яскраві. Серед них такі як спрага, часті позиви до туалету, нечіткий зір, погіршення настрою, поколювання в руках, постійна втома, погане загоєння ран, розповсюдження інфекційної хвороби.

Діабет 2 типу може розвинутися в будь-якому віці, навіть у дитинстві. Однак цукровий діабет 2 типу найчастіше зустрічається у людей середнього та старшого віку. Більша ймовірність розвитку діабету 2 типу, якщо людині 45 років або старше, є сімейна історія діабету, надмірна вага чи ожиріння. Діабет частіше зустрічається у афроамериканців, латиноамериканців, американських індіанців, азіатських американців або жителів тихоокеанських островів.

Відсутність фізичної активності та певні проблеми зі здоров’ям, такі як високий кров’яний тиск, впливають на ймовірність розвитку діабету 2 типу. Більша ймовірність розвитку діабету 2 типу, якщо є переддіабет або був

гестаційний діабет під час вагітності.

Лікування діабету – не одноразова терапія, а процес на усе життя. Діабет потребує великої уваги до здоров'я, особливо що стосується моніторингу життєвих показників:

- контроль рівня глюкози в крові
- контроль артеріального тиску
- догляд за судинами великого калібру
- скринінг і лікування ретинопатії (яка спричиняє сліпоту)
- контроль ліпідів крові (для регулювання рівня холестерину)
- скринінг на ранні ознаки захворювання нирок, пов'язаного з діабетом, та лікуванням.

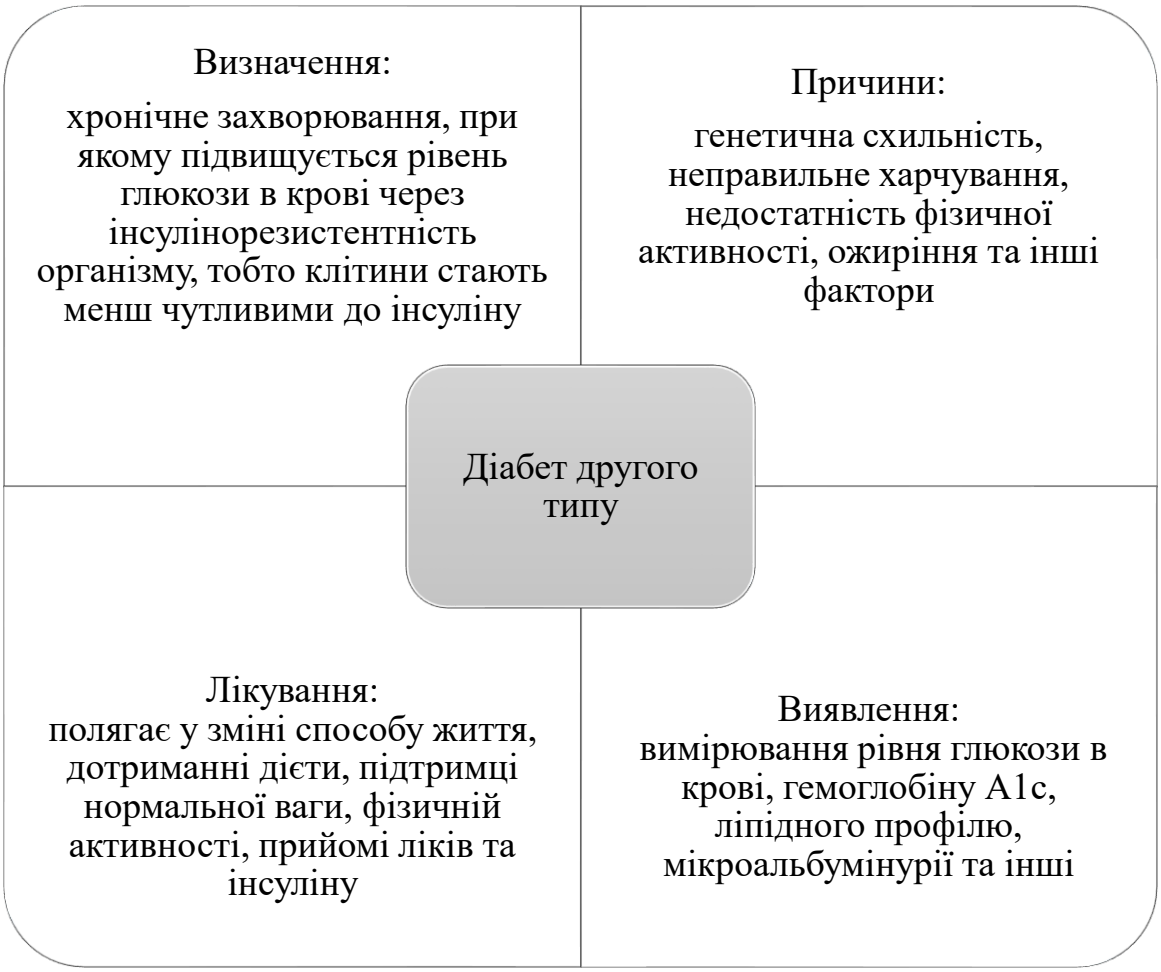

Рис.1.1. Зведена характеристика про захворювання ДДТ*

*авторська розробка

Дані 10-го видання IDF Diabetes Atlas повідомляють про постійне

глобальне зростання поширеності цукрового діабету, що підтверджує, що діабет є серйозною глобальною проблемою для здоров'я та благополуччя окремих людей, сімей і суспільства.

537 мільйонів дорослих (20-79 років) живуть з цукровим діабетом – це означає, що діабетом хворіє кожен 1 з 10. За прогнозами, до 2030 року ця цифра зросте до 643 мільйонів і до 784 мільйонів у 2045 році. Діабет є причиною 6,7 мільйонів смертей у 2021 році – тобто 1 людина помирала кожні 5 секунд у зв'язку з діабетом [24,20,52,4,1,9,3].

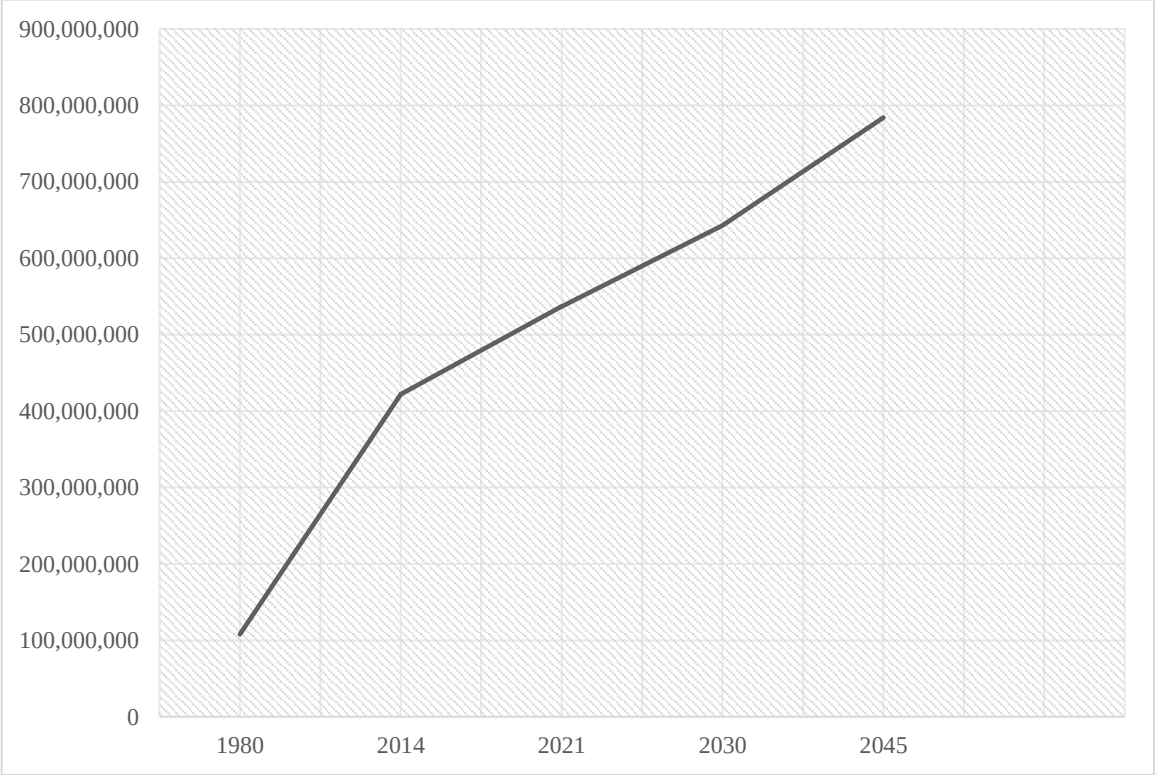

Рис. 1.2. Кількість хворих на діабет за даними IDF Diabetes Atlas та WHO *[29]

Сьогодні в Україні 1 мільйон 134 тисячі хворих на цукровий діабет (станом на 2020 рік), проте точно наразі неможливо дослідити тенденції захворюваності на цукровий діабет через відсутність статистики щодо смертності внаслідок цієї хвороби [73].

Своєчасне виявлення діабету вкрай покращує його лікування. Процес виявлення цукрового діабету є складним на ранній стадії, особливо в Україні. В Україні невеликий відсоток людей робить так звані check-up, перевірку усього

організму за рахунок аналізів. На те є декілька причин: недовіра до медицини, низький рівень зарплатні, низький рівень освіти в питаннях здоров'я. Таким чином люди з діагнозом діабет можуть не здогадуватися про свій стан та потрапити до лікарні із тяжким станом через супутні захворювання, інсульт чи інфаркт. Водночас, навіть наявність ряду аналізів не гарантує відгуку лікарів на проблему, і постановка діагнозу залишається в руках пацієнта.

Розвиток підходів до машинного навчання здається вирішує цю проблему. Маючи результати аналізів можна визначити стадію та вірогідність діагнозу застосувавши один із методів Data Science – машинне навчання. Водночас відкритим залишається питання того, як вибрати дійсні ознаки хвороби та правильний класифікатор.

Постає потреба в особливому продукту – застосунку для передбачення діабету за допомогою алгоритмів машинного навчання.

Дана потреба є стабільною, про що говорять Google Trends за ключовим запитом «diabetes symptoms». Тобто люди постійно цікавляться чи не хворіють вони на діабет, чи мають виражені симптоми, проте не мають інструменту надійно отримати прогноз. Вони інтерпретують симптоми на власний розсуд, не йдуть ані до лікарів, ні до лабораторій.

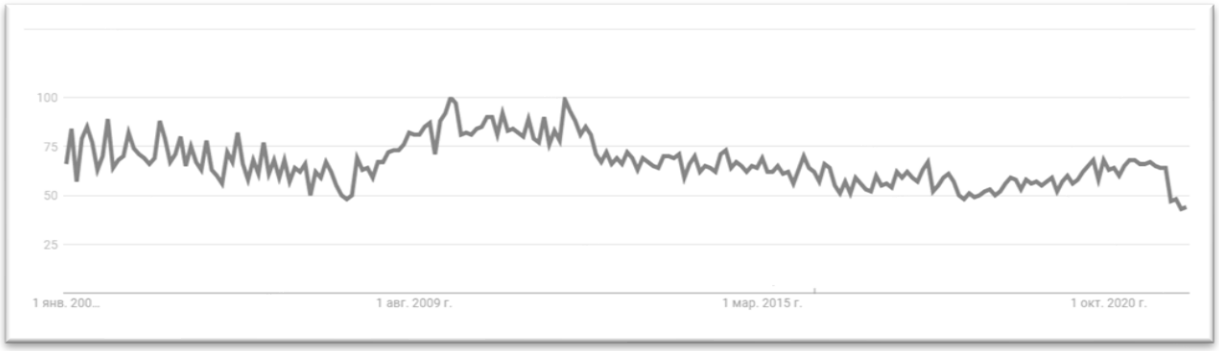

Рис. 1.3. Динаміка запитів щодо симптомів діабету*

*[68]

Маючи дієвий інструмент, який користується надійною базою даних пацієнтів та їх аналізів, людина може здати необхідні аналізи та отримати результат прогнозу методами машинного навчання.

Сформулюємо дерево проблем та дерево рішень для даного дослідження.

- Недостатнє збирання інформації про пацієнтів з діабетом другого типу
    - Недостатньо даних про симптоми, лікування та попередження хвороби
    - Відсутність централізованої бази даних для зберігання інформації про пацієнтів та їхніх лікарів
- Неефективне використання даних про пацієнтів з діабетом другого типу
    - Недостатнє використання аналітики даних для знаходження та аналізу трендів та паттернів у здоров'ї пацієнтів
    - Відсутність засобів для швидкого та ефективного спілкування між пацієнтами та їхніми лікарями
- Низька свідомість пацієнтів про ДДТ та можливі наслідки хвороби
    - Відсутність ресурсів та засобів для налагодження освіти та просвітницьких кампаній про ДДТ
    - Відсутність доступної та зрозумілої інформації для пацієнтів про симптоми, лікування та профілактику хвороби.

Аналізуючи дерево проблем можна сформулювати задачу «створення технології визначення діабету в зручній формі», яку можна розвинути до дерева рішень і опису потенційного рішення проблематики.

Рис.1.4. Дерево проблем та дерево рішень*

*авторська розробка

Дерево рішень:

- Використання машинного навчання для діагностики діабету може допомогти зменшити кількість помилкових діагнозів, що може знизити ризик неправильного лікування та погіршення стану пацієнтів з діабетом другого типу.

- Розробка алгоритмів машинного навчання, що зможуть ефективно класифікувати дані тестів на діабет, є ключовим фактором успіху проекту.
- Навчання алгоритмів машинного навчання на достатньо великій та репрезентативній вибірці даних може покращити точність діагностики та зменшити кількість помилкових діагнозів.

- Розробка застосунку для збору даних та діагностики може забезпечити зручний та доступний інструмент для використання пацієнтами та медичним персоналом.
  - Розробка інтерфейсу, який би дозволяв зручно збирати дані від пацієнтів
  - Забезпечення безпеки та конфіденційності даних, що збираються в мобільному застосунку
- Розробка бази даних для зберігання та обробки результатів тестів може допомогти зберегти та організувати інформацію, необхідну для діагностики та лікування ДДТ.
  - Розробка структури бази даних, яка би забезпечувала ефективне зберігання та організацію даних
  - Розробка механізмів аналізу та обробки даних

1.2. Вирішення проблеми визначення захворювань в медичній сфері методами машинного навчання на підприємствах

У сучасному світі медична сфера постійно стикається з викликами, пов'язаними з діагностикою та визначенням захворювань. Вирішення цих проблем є вкрай важливим, оскільки швидка та точна діагностика є критично важливою для ефективного лікування та збереження здоров'я пацієнтів.

Одним з інноваційних підходів до визначення захворювань є використання методів машинного навчання на підприємствах медичної сфери. Машинне навчання відіграє ключову роль у вирішенні проблеми точності та швидкості

діагностики, завдяки своїм здатностям аналізувати великі обсяги медичних даних та виявляти складні взаємозв'язки.

На підприємствах використання методів машинного навчання для прогнозування та діагностики діабету може мати позитивний вплив на ефективність та якість медичної допомоги, а також знизити витрати на охорону здоров'я. Раннє виявлення та ефективне управління діабетом може зменшити кількість госпіталізацій та ускладнень, що може бути корисно як для пацієнтів, так і для підприємств, що надають медичні послуги.

Застосування методів машинного навчання на підприємствах медичної сфери відкриває широкий спектр можливостей. Вони дозволяють автоматизувати процес діагностики, прогнозування та класифікації захворювань, що призводить до зниження витрат часу та зусиль медичних фахівців. Крім того, вони дозволяють отримувати більш точні результати, оскільки машинне навчання враховує багатофакторність та складні взаємозв'язки між різними показниками здоров'я.

Хоча всі ці переваги й очевидні, такі технології потребують як спеціалістів так і інвестицій. Наразі в Україні немає подібних систем. Проте вони є частково впровадженні в світі у різних проектах.

«Cardiogram by ilumivu»: використовує DeepHeart – глибоку нейронну мережу для прогнозування серцево-судинних захворювань на основі зібраних даних [8].

«One Drop»: це мобільний додаток, який дозволяє користувачам з діабетом вести контроль над своїм станом за допомогою моніторингу рівня цукру в крові, внесенням даних про прийом їжі, фізичні вправи та інші фактори. Система використовує алгоритми машинного навчання для надання рекомендацій щодо дозування інсуліну та інші поради для керування діабетом [34].

«Sweetch»: це ізраїльський стартап, що розробляє систему моніторингу рівня цукру в крові та вивчення звичок користувачів для визначення ризику розвитку діабету. Система використовує аналіз даних за допомогою машинного навчання для надання рекомендацій щодо життєвого стилю та дієти [54].

«DreaMed Diabetes»: ця система, розроблена в Ізраїлі, використовує алгоритми машинного навчання для розрахунку індивідуальних доз інсуліну для пацієнтів з діабетом. Система збирає дані про рівень цукру в крові, прийом їжі та інші фактори і надає рекомендації [21].

Нижче наведена порівняльна таблиця з перевагами та недоліками чотирьох систем.

Таблиця 1.1

Порівняльна характеристика конкурентів*

Назва системи	**Переваги**	**Недоліки**
Cardiogram by ilumivu	Використовує глибоке навчання для прогнозування ризику розвитку діабету та інших захворювань серцево-судинної системи; Можливість моніторингу за допомогою носимих пристроїв; Використання технології машинного навчання для навчання системи на основі нових даних.	Потребує використання спеціальних пристроїв для моніторингу; Висока ціна використання системи.
One Drop	Застосування машинного навчання для прогнозування рівня глюкози в крові та інших показників здоров'я; Можливість моніторингу за допомогою смартфонів та інших мобільних пристроїв; Використання штучного інтелекту для підбору індивідуальної дієти та рекомендацій по контролю за здоров'ям.	Висока ціна використання системи; Потребує використання спеціальних пристроїв для моніторингу.
Sweetch	Підбір індивідуальних рекомендацій щодо здорового способу життя.	Обмежені можливості моніторингу (не вимірює рівень глюкози в крові); Висока ціна використання системи.
DreaMed Diabetes	Система має високу точність визначення необхідної кількості інсуліну, що допомагає пацієнтам контролювати рівень глюкози в крові та запобігати ускладненням	Система поки не доступна для широкого споживача і використовується головним чином в медичних установах

*авторська розробка

Як бачимо закордоном більш розвинута дана сфера, проте також немає цілісної системи для лікарів і здебільшого орієнтація на виключні функції.

Не зважаючи на те, що в Україні подібних систем немає, в нас широко використовується машинне навчання для інших сфер.

Так на прикладі ТОВ Інфопульс, з використанням методів машинного навчання.

Найпоширенішими проектами з використання методів машинного навчання є проекти виявлення аномалій. У загальному це системи, що використовують розвідку в реальному часі на основі алгоритмів машинного навчання для створення математичних моделей виявлення аномалій, спеціально підібраних для кожного набору даних. Система забезпечує зручну візуалізацію даних через панелі керування та аналітичні звіти з детальними практичними висновками. Відгуки користувачів використовуються для поліпшення рішення.

Першим прикладом такої системи є проект компанії Infopulse, в якому використовувалися ММН для розробки системи моніторингу газових турбін, що дозволяє виявляти аномальну роботу турбін та надавати операторам попередження про можливі поломки.

Проект був реалізований для партнера компанії Infopulse – виробника газових турбін. Основною метою проекту було зменшення витрат на планове обслуговування газових турбін та зниження ризику їх виходу з ладу внаслідок неочікуваних проблем.

У проекті було використано ММН, зокрема класифікацію та кластеризацію даних, для розробки математичних моделей, які аналізували дані з сенсорів газових турбін. На основі цих моделей було розроблено систему моніторингу газових турбін, що забезпечує раннє виявлення аномальної роботи турбін.

Система моніторингу була інтегрована з веб-інтерфейсом, що надавав операторам доступ до панелі управління. Через цей інтерфейс оператори могли отримувати оперативну інформацію про стан газових турбін та отримувати попередження про можливі поломки.

Результатом проекту стало значне зменшення витрат на планове

обслуговування газових турбін та зниження ризику їх виходу з ладу внаслідок неочікуваних проблем. Крім того, система моніторингу була успішно впроваджена в роботу, що дозволило покращити ефективність виробництва та знизити витрати на ремонт газових турбін [27].

Наступним прикладом є використання згорткових нейронних мереж (Convolutional Neural Network, CNN) для задачі визначення кількості елементів на виробничій лінії. Компанія мала проблему з ручним підрахунком кількості елементів, що проходять через виробничу лінію, що забирає багато часу та може бути неточним.

Щоб вирішити цю проблему, Infopulse розробив CNN модель, яка отримує зображення з камер виробничої лінії та визначає кількість елементів на зображенні. Модель була тренована на наборі даних, що складався з тисяч зображень елементів на виробничій лінії. Для цього було використано бібліотеку Keras.

Після тренування моделі вона була впроваджена на виробничу лінію, де вона автоматично зчитує зображення та визначає кількість елементів. Інтеграція моделі з виробничою лінією забезпечує точність визначення кількості елементів та зменшує час, що потрібен для підрахунку. Крім того, система збирає дані про виробничий процес, що дозволяє проводити аналіз та оптимізувати продуктивність виробничої лінії [12].

Також є приклад проекту в агрокомплексі [46].

У цьому кейсі компанія Infopulse працювала з клієнтом, який виготовляє борошняні вироби. Метою проекту було зменшення відхилень якості продукції, зниження витрат на виробництво та оптимізація процесів контролю якості.

Для цього Infopulse створила прогнозуючу модель якості, що базувалась на аналізі великого обсягу даних про якість продукції та виробничих процесів. У модель включалися такі фактори, як тип сировини, властивості виробництва та вплив погодних умов.

Далі Infopulse розробила систему моніторингу якості, яка включала в себе відстеження параметрів виробництва в реальному часі та порівняння їх з

прогнозом моделі. При виявленні відхилень система автоматично сповіщала відповідних фахівців, які займались контролем якості, щоб вони могли вчасно прийняти заходи для усунення проблем.

Результатом проекту стало значне зниження кількості відхилень якості продукції та оптимізація витрат на виробництво. Також система моніторингу якості дозволяє в реальному часі відстежувати процеси виробництва та приймати оперативні рішення щодо їх оптимізації.

Технології даних проектів різноманітні і залежать як від технічного складу команди розробників, так і від специфічних побажань клієнтів. Але у загальному це Python, PyTorch, Theano, TensorFlow, Keras, Spark, Azure Machine Learning.

Таким чином, машинне навчання стає все більш популярним серед українського IT та вже зараз надаються послуги компаніями на прикладі Інфопульс, оскільки воно дозволяє покращувати наявні рішення в багатьох сферах життя.

Рис.1.5. Використання машинного навчання у різних сферах людської діяльності*

*авторська розробка

Однією з найбільш потребуючих машинного навчання сфер є медицина. Машинне навчання допомагає аналізувати медичні зображення, виявляти аномалії і діагностувати захворювання. Також воно може допомогти у плануванні терапії та лікуванні пацієнтів використовуючи медичні дані.

Інші сфери, які користуються машинним навчанням, включають:

Фінанси і банківський сектор: машинне навчання використовують для виявлення шахрайства, ризик-аналізу та прогнозування ринків.

Транспорт: машинне навчання використовують для вдосконалення систем автоматичного керування, безпеки на дорогах та оптимізації маршрутів.

Реклама та маркетинг: машинне навчання використовують для персоналізованої реклами, рекомендацій продуктів та побудови різних моделей прогнозування.

Виробництво: машинне навчання може використовуватися для контролю якості продуктів, оптимізації процесів виробництва, виявлення аномалій та передбачення відмов обладнання.

Сфера безпеки: машинне навчання може використовуватися для розпізнавання облич, відслідковування та аналізу відеозаписів, виявлення загроз та інших завдань, що пов'язані з безпекою.

1.3. Аналіз необхідності застосування методів машинного навчання для прогнозування ДДТ

Застосування методів машинного навчання для прогнозування ДДТ є необхідним, оскільки це дає можливість створити моделі, які можуть передбачати ризик розвитку цієї хвороби в зазначеного пацієнта. Такі моделі можуть використовуватися лікарями, щоб проводити скринінг ризику діабету у пацієнтів, які не мають симптомів цієї хвороби або знаходяться на початкових стадіях.

Застосування методів машинного навчання також дозволяє точніше прогнозувати, які фактори можуть впливати на розвиток діабету, такі як вік,

стать, історія сімейного захворювання, рівень активності, дієта та інші фактори. Завдяки цьому лікарі можуть забезпечити більш індивідуалізоване лікування та попередження розвитку хвороби.

Крім того, застосування методів машинного навчання дозволяє зменшити кількість помилкових діагнозів та забезпечити більш точну діагностику діабету, що є особливо важливим при плануванні лікування та попередженні ускладнень.

Використання машинного навчання в медицині може бути необхідно для багатьох імплементацій. Одним із таких впроваджень є покращення сервісу у наданні медичних послуг а також покращення виявлення та лікування захворювань.

Інше основне застосування алгоритмів машинного навчання в охороні здоров’я полягає в тому, що можна виявляти проблеми зі здоров’ям до того, як вони стануть хронічними хворобами, аналізуючи масивні дані пацієнтів. Клінічні установи можуть застосовувати алгоритми машинного навчання для виявлення інсультів на основі існуючих станів здоров’я, оцінки здоров’я серця та виявлення інших проблем.

Лікарі та клінічні експерти можуть діагностувати захворювання набагато раніше, оскільки алгоритми забезпечують аналіз у реальному часі [51].

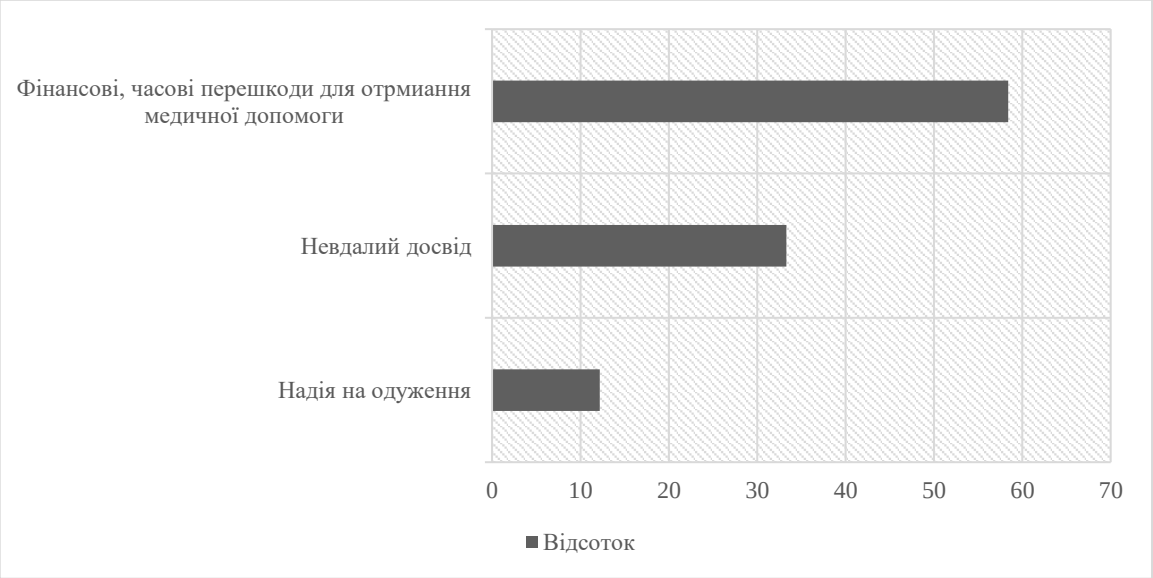

Рис.1.6. Розподіл причин відмови від звернення за медичною допомогою* *[56]

Дослідження показують, що люди уникають звернень за медичною допомогою, навіть якщо підозрюють певний діагноз. Визначено три основні категорії причин ухилення від медичної допомоги. По-перше, понад третина учасників (33,3% з 1369) повідомили про поганий досвід звернення за медичною допомогою, наприклад, не чемні лікарі, погано організовані організації охорони здоров’я та емоційні проблеми пов’язані з розголосом лікарської таємниці. По-друге, підгрупа учасників повідомила про низьку потребу звернутися за медичною допомогою (12,2%), часто тому, що вони очікували, що їхня хвороба або симптоми покращяться з часом (4,0%). По-третє, багато учасників повідомили про традиційні перешкоди на шляху до медичної допомоги (58,4%), такі як висока вартість (24,1%), відсутність медичного страхування (8,3%) та часові обмеження (15,6%).

Для того, щоб встановити діабет недостатньо одного аналізу, що означає тривалість даного процесу для пацієнта і можливість отримати окрім діагноза ще й психологічну травму. Маючи ж пакет аналізів та застосунок, який інтерпретує ці дані, можна зрозуміти, який в пацієнта статус на даний момент. Також маючи щорічні аналізи можна побачити динаміку та зробити висновки завчасно.

Водночас впровадження машинного навчання до медицини має певні виклики. Перша проблема полягає в тому, що медичні дані зазвичай містять велику кількість пропущених значень, неповних записів та інші невідомі значення. Це може вплинути на точність та якість моделі, яку буде створено на основі цих даних.

Друга проблема пов'язана з обмеженими ресурсами для збору та підготовки медичних даних. Для створення ефективної моделі необхідно мати достатньо великий обсяг даних, які повинні бути коректно зібрані, підготовлені та очищені від помилок та шумів. Однак, в більшості випадків, доступні обмежені обсяги даних, або вони мають низьку якість, що ускладнює створення моделі.

Третя проблема пов'язана з конфіденційністю медичних даних. Оскільки медичні дані містять особисту інформацію про пацієнтів, то виникає проблема

забезпечення їх конфіденційності. Для розв'язання цієї проблеми необхідно застосовувати спеціальні методи захисту персональних даних.

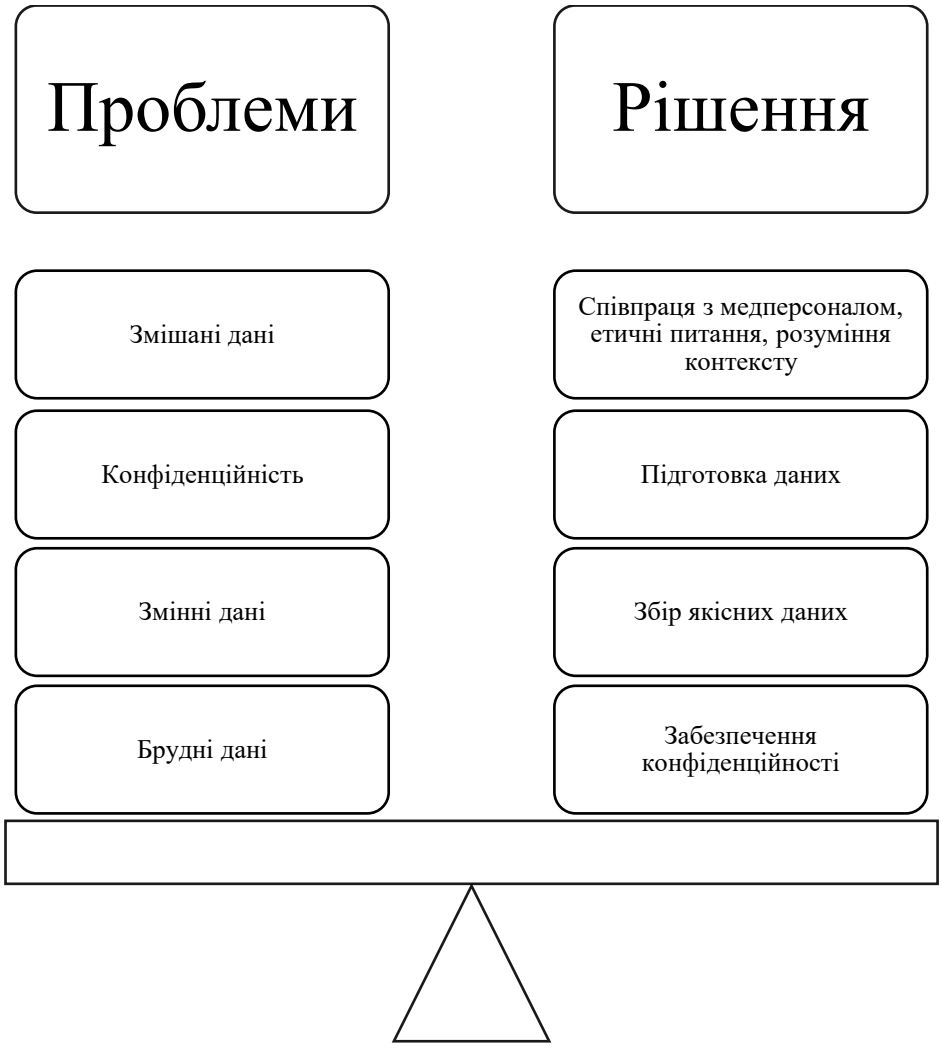

Рис.1.7. Баланс між ризиками та рішеннями питання впровадження машинного навчання до медичної сфери*

*авторська розробка

Четверта проблема пов'язана зі змінністю медичних даних. Медичні дані зазвичай мають складну структуру та можуть містити різноманітні дані, які змінюються з часом. Це може ускладнити створення стійкої моделі, яка буде коректно працювати з даними, які змінюються з часом.

Остання проблема пов'язана зі змішаними даними. Медичні дані можуть містити як числові, так і категоріальні дані. При створенні моделі необхідно враховувати наявність різноманітності в даних.

1.4. Висновки до першого розділу

У першому розділі дослідження було детально розглянуто предметну область захворювання ДДТ та його визначення. З метою збільшення розуміння та глибшого аналізу проблеми, була надана детальна характеристика захворювання, що охоплювала як його причини, так і можливості виявлення. Важливою складовою цього дослідження було визначення трендів розвитку захворювання та попиту у населення стосовно цього питання. Було з’ясовано, що ДДТ стає все більш поширеним у світі та є одним з найбільш впливових захворювань на здоров’я населення.

Далі, була проведена аналітична робота зі світової та української практики визначення ДДТ, під час якої було розглянуто декілька рішень та знайдено їх переваги та недоліки. Було виявлено, що на сьогоднішній день в Україні немає конкурентоспроможного рішення для визначення ДДТ, тому потреба в такому рішенні є нагальною.

Після цього, було з’ясовано, що одного лікарського висновку не є достатньо для точного та надійного визначення ДДТ. Тому, для підвищення ефективності визначення цього захворювання, було запропоновано використати ММН. Ці методи дозволяють зібрати та обробити великі обсяги даних, що дозволяє забезпечити точність та надійність діагностики.

Отже, була поставлена задача для вирішення проблеми низького та неточного визначення ДДТ у створенні технології визначення за допомогою машинного навчання.

РОЗДІЛ 2. МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ ДІАБЕТУ ДРУГОГО ТИПУ

2.1. Аналіз алгоритмів машинного навчання

У статті Forbes, яка з’явилася в лютому 2018 року, машинне навчання поміщено на другу позицію в рейтингу найвпливовіших технологій найближчого майбутнього.

ММН поділяють на три групи: навчання з вчителем, навчання без вчителя, навчання з підкріпленням.

У випадку машинного навчання з учителем в машині є «наставник», який підказує їй, як правильно виконувати свою задачу. Учитель заздалегідь зазначає всі необхідні дані, щоб машина могла засвоїти інформацію на конкретних прикладах. Машина набагато краще і швидше вчиться разом з учителем, тому такі алгоритми частіше використовуються для вирішення практичних завдань. Алгоритми навчання з учителем включають такі види задач, як регресія та класифікація.

Модуль регресійного аналізу зазвичай використовується для того, щоб передбачити зв’язок між процесом і результатом. Такий аналіз може показати, як змінні пов’язані одна з одною, а визначення причин і наслідків є предметом більш глибокого дослідження з використанням інших алгоритмів і методів. Графік регресії може показувати позитивний зв’язок, негативний зв’язок або відсутність зв’язку з певними факторами. Якщо лінія регресії горизонтальна або вертикальна, між змінними немає зв’язку.

Алгоритми класифікації дозволяють розділити об'єкти за заздалегідь заданими класами. Сьогодні алгоритми класифікації використовуються для великої кількості завдань: виявлення мови, спам-фільтри, виявлення шахрайства, виявлення захворювань. Щоб класифікація працювала, потрібно мати позначені дані з категоріями та характеристиками, які машина навчиться ідентифікувати. Залежно від тих чи інших ознак алгоритм визначає, до якого з класів можна

віднести об’єкт.

На тренуваннях машини без викладача машина як зрозуміло навчається без підказок. Дані в цьому випадку не позначені певними категоріями, а машина сама намагається знайти закономірності. На практиці такі алгоритми використовуються рідше, як правило, як методи аналізу та підготовки даних, а не як базовий алгоритм, що вирішує конкретні задачі з використанням цих даних.

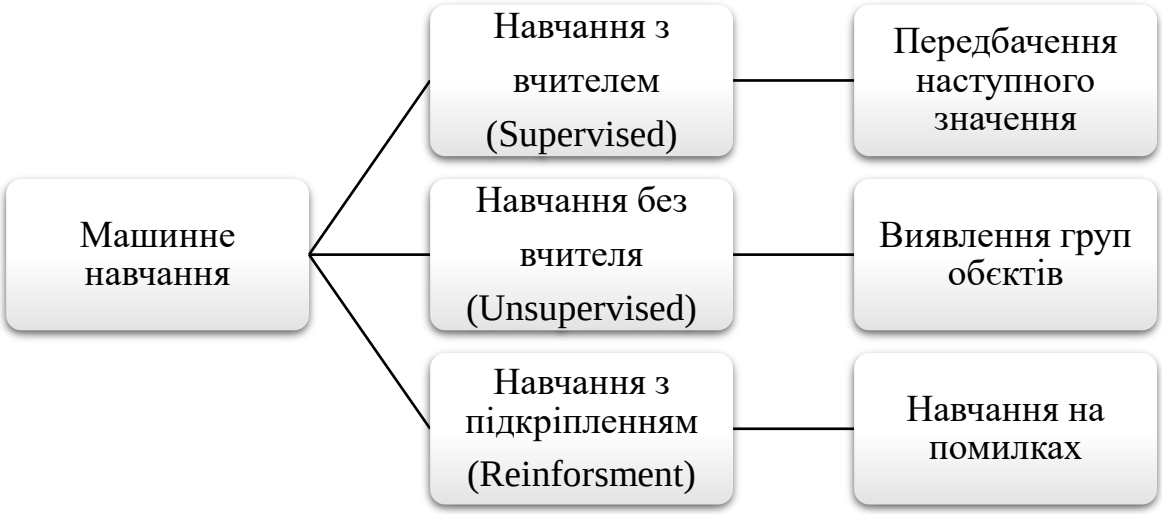

Рис.2.1. Типи машинного навчання*

*[28]

Посилене навчання менш схоже на попередні типи, оскільки воно більше нагадує штучний інтелект. Такі алгоритми використовуються не в аналізі даних, а в реальному середовищі. Наприклад, автопілот Tesla, який в симуляції вчиться не збивати пішоходів. Завдання таких машин — не прорахувати всі ходи, а мінімізувати помилки або максимізувати прибуток. Посилене навчання дуже схоже на власне навчання людей - автомобіль карається за помилки і заохочується за правильні дії.

Кероване навчання є найпопулярнішою парадигмою машинного навчання. Він найпростіший для розуміння і найпростіший у реалізації.

Найтиповішими прикладами керованого навчання є:

- Маркетинг: вибір реклами, яка буде ефективною, алгоритми визначають ті оголошення, які будуть мати найбільші охоплення, за рахунок дослідження поведінки користувачів в минулі періоди.
- Класифікація спаму: фільтр спаму розпвізнає мітки (спам/не спам), завчасно відфільтровувавши шкідливі електронні листи.

Навчання без нагляду є протилежним навчанню з наглядом. В алгоритм надається багато даних та інструменти для розуміння властивостей даних. Звідти він може навчитися групувати, та/або організовувати дані таким чином, щоб людина могла увійти і зрозуміти щойно організовані дані.

Серед прикладів таких методів можна навести:

- Системи рекомендацій: YouTube, Netflix
- Купівельні звички: купівельні звички містяться у базі даних, і ці дані з активно купуються та продаються. Ці купівельні звички можна використовувати в алгоритмах навчання без нагляду для групування клієнтів у подібні сегменти покупок. Це допомагає компаніям виходити на ринок цих згрупованих сегментів і може навіть нагадувати рекомендаційні системи.

Навчання з підкріпленням досить відрізняється від навчання під наглядом і навчання без нагляду. Прикладами навчання з підкріпленням є:

- Відеоігри: AlphaZero та AlphaGo, які навчилися грати в гру Go.
- Промислове моделювання: для багатьох роботизованих додатків корисно, щоб машини навчалися виконувати свої завдання без необхідності жорсткого кодування своїх процесів.

Серед популярних методів машинного навчання виділяють наступні: Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), дерево рішень (DT), логістисна регресія (LR), Random Forest (RF).

2.2. Вирішення проблеми визначення діабету методами машинного навчання

Використання методів машинного навчання є гострим питанням і активно досліджується. Одними із останніх досліджень в цій сфері є:

У 2022 році було проведено дослідження авторів Poly TN, Islam MM та Li YJ щодо можливості передбачення діабету 2 типу з використанням методів машинного навчання. Дослідники порівняли ефективність трьох різних методів машинного навчання: методу опорних векторів, дерева рішень та логістичної регресії.

Для експерименту використовувалися дані пацієнтів, які мали або не мали діабет, зібрані з медичних записів. Результати дослідження показали, що метод опорних векторів та дерево рішень мають найкращу ефективність у передбаченні діабету, з точністю близько 80%, тоді як логістична регресія мала меншу ефективність з точністю близько 70% [36].

Також у 2022 році Dutta et al. провели дослідження ранньої діагностики діабету, використовуючи ансамбль моделей машинного навчання. Автори зібрали клінічні дані про пацієнтів з діабетом другого типу та здорових осіб з метою побудови моделі для передбачення ризику виникнення діабету [22].

Автори використовували такі ММН, як логістична регресія, дерево рішень, а також ансамбль моделей, такий як випадковий ліс та градієнтний бустинг. Для оцінки точності моделей використовувалися метрики, такі як точність, чутливість та специфічність.

Результати дослідження показали, що ансамбль моделей машинного навчання має кращу точність та чутливість в порівнянні з окремими моделями.

У дослідженні Deberneh HM та Kim I було проведено прогнозування ДДТ за допомогою алгоритмів машинного навчання на наборі даних з клінічної інформації про пацієнтів, який містив такі характеристики, як вік, стать, рівень глюкози, індекс маси тіла та інші [14].

Було проведено порівняльний аналіз різних моделей машинного навчання для прогнозування діабету. В результаті дослідження було виявлено, що модель логістичної регресії має найвищу точність порівняно з іншими алгоритмами машинного навчання, такими як метод опорних векторів та дерево рішень.

У роботі G. Tripathi and R. Kumar було досліджено декілька алгоритмів машинного навчання для класифікації діабету, зокрема логістичну регресію, дерево рішень, SVM та Random Forest.

Дослідники зібрали дані від 768 пацієнтів, використовуючи вісім ознак, таких як вік, BMI, кількість вагітностей тощо, та розподілили їх на два класи: діабет і недіабет. Для порівняння різних алгоритмів використовувалися метрики точності, чутливості та специфічності.

Дослідження показало, що Random Forest був найкращим алгоритмом для класифікації діабету, з точністю 97,66%, чутливістю 96,52% та специфічністю 98,03%. Дерево рішень та SVM також показали добрі результати з точністю більше 90%, але Random Forest перевершив їх у всіх метриках [26].

Talha Mahboob Alam та інші підтвердили ефективність використання Random Forest для діагностики діабету та показали, що розроблена система може бути корисною для медичних працівників у процесі визначення діагнозу та розробки плану лікування [57].

Shamreen Ahamed, Meenakshi Sumeet Arya та Auxilia Osvin Nancy V у статті 2022 року у журналі Frontiers in Computer Science визначили, що «алгоритм LGBM» найкраще працював для взятого набору даних, забезпечуючи вищу точність порівняно з іншими алгоритмами. Також вони визначили, що з різними наборами даних, різні класифікатори дають кращі результати. Крім того, параметри, які використовуються в LGBM, можуть бути додатково налаштовані, можна використовувати розширений алгоритм LGBM щоб збільшити відсоток точності передбачення [43].

C. Charitha у стаття "Автоматичне визначення рівня цукру в крові за допомогою машинного навчання на основі методів аналізу даних" описує ретельну попередню обробку даних і розробку функцій за допомогою моделей

важливості функцій, таких як Random Forest Importance і RFE, автори використали багато моделей машинного навчання, таких як KNN, Logistic Regression, SVM, Random Forest, LightGBM і XGBoost для прогнозування цукрового діабету II типу. Серед усіх моделей найвищу точність отримано в 91,47% від моделі lightGBM. [61].

У статті «Прогнозування діабету 2 типу за допомогою методів класифікації машинного навчання» 2020 року автори Neha Prerna Tigga та Shruti Garg було реалізовано шість методів класифікації машинного навчання, а їх результати порівнювалися з різними статистичними показниками. Тести проводили на наборі даних, зібраних за допомогою онлайн- і офлайн-анкет, що складалися з 18 запитань, що стосуються діабету. Також ті самі алгоритми були застосовані до бази даних PIMA. Результат експерименту показує, що точність Random Forest набору даних становить 94,10%, що є найвищим показником серед інших. Випадковий ліс також забезпечує найвищу точність для набору даних PIMA. Серед шести різних застосованих алгоритмів машинного навчання всі моделі дали хороші результати для деяких параметрів, таких як точність, чутливість до запам'ятовування тощо. Було визначено, що «Вік», «Сімейний діабет», «Фізична активність», «Звичайне лікування» та «Діабет» або гестаційний діабет має найвищу значимість серед усіх змінних. Ці параметри мають більший вплив на прогноз діабету, ніж інші [42].

У дослідженні Parisa Karimi Darabi, Mohammad Jafar Tarokh було досліджено декілька методів машинного навчання та встановлено, що для набору даних з Китаю найкраще підійшов метод градієнтового бустингу [59].

У дослідженні 2019 року від V. Vaidehi до набору даних застосовано різні алгоритми машинного навчання, а класифікація виконана за допомогою різних алгоритмів, з яких логістична регресія дає найвищу точність 96% [17].

Таким чином було представлено зведену інформація про дослідження, присвячені розробці алгоритмів машинного навчання для прогнозування діабету. Дослідження були проведені у 2020-2022 роках і включали в себе використання різних алгоритмів машинного навчання, таких як Random Forest, SVM, Naive

Bayes, Decision Tree, XGBoost і LightGBM, для класифікації діабету. В загальному дослідження показують що алгоритм має підбиратися під конкретний набір даних, тестуватися та можливо поєднуватися у ансамбль для покращення результатів. ММН дають високу точність для прогнозування ДДТ.

У сфері визначення діабету використовуються різні алгоритми машинного навчання, такі як логістична регресія, дерево рішень, нейронні мережі, метод опорних векторів (SVM) та інші. Кожен з цих алгоритмів має свої переваги та недоліки, і вибір алгоритму залежить від конкретної задачі та вхідних даних.

ММН використовується доволі давно, приблизно з 2008 року у роботах дослідників прогнозування цукрового діабету. Водночас вони не втрачають своєї актуальності, адже використовуються різні параметри, методи та джерела даних.

Метод логістичної регресії є одним з методів машинного навчання, що використовується для класифікації даних. Він дозволяє передбачити категорію вихідного сигналу (наприклад, чи клієнт купить продукт чи ні) з використанням набору вхідних даних (наприклад, вік, стать, дохід тощо).

Математично, логістична регресія базується на функції логістичної кривої, яка перетворює ваговану суму вхідних даних на імовірність належності до певного класу.

У статті 2021 року Ram D. Joshi1,та Chandra K. Dhakal на тему «Прогнозування діабету 2 типу за допомогою підходів логістичної регресії та машинного навчання» було досліджено метод логістичної регресії та досягнуто результатів: виявлено п’ять основних прогностичних факторів діабету 2 типу: рівень глюкози, вагітність, індекс маси тіла (ІМТ), родові функції та вік. Створено модель із точністю 78,26% і рівень помилок 21,74%. Автори стверджують, що їх модель може бути застосована для обґрунтованого прогнозування діабету 2 типу та потенційно може бути використана для доповнення існуючих профілактичних заходів для стримування захворюваності на діабет та зменшення пов’язаних з цим витрат [39].

У статті 2021 року від Priyanka Rajendra була розроблена модель із логістичною регресією, було відібрано основні незалежні змінні, було досягнуто

точності залежно від набору даних від 74 до 88% [41].

Дерево рішень (Decision tree) - це метод машинного навчання, який можна використовувати для класифікації та регресії. Він працює шляхом створення дерева рішень зі списком правил, які можуть бути використані для прийняття рішень. Кожна гілка дерева представляє можливу варіацію дій, які можуть бути прийняті відповідно до певного критерію розбиття. При вирішенні задачі класифікації дерево рішень розбиває навчальні дані на підмножини, де кожна підмножина відноситься до конкретного класу, тобто містить лише предмети цього класу.

У дослідження 2019 року у статті «Аналіз дерева рішень для прогнозування діабету» автори досліджували різні алгоритми на основі дерев і визначили, що їх точність з вибраним набором даних коливається віл 88 до 95,8% [5].

Нейронна мережа (Neural network) - це алгоритм машинного навчання, який створюється на основі імітації роботи людського мозку. Вона складається з нейронів, які співпрацюють між собою та передають інформацію через зв'язки між ними. Нейронна мережа може використовуватися для класифікації, регресії та інших задач обробки даних. Під час тренування нейронної мережі, вона навчається розпізнавати зв'язки між вхідними даними та вихідними результатами.

У статті 2020 року під назвою «Тип 2: Прогнозування цукрового діабету за допомогою класифікатора Deep Neural Networks» було досягнуто таких висновків, як запропонована DNN-FI модель отримала кращий рівень точності порівняно з алгоритмом випадкового лісу та дерева рішень. Ця модель досягла 96,77% у коефіцієнті розподілу тестів 60–40, 97,54% у коефіцієнті розподілу тестів 70–30, 98,16% у коефіцієнті розподілу тестів 80–20 і 96,10% у 10-кратній перехресній перевірці. Основним обмеженням методу є обчислювальний час [60].

Робота інших авторів за схожою тематикою «Модель штучних нейронних мереж для прогнозування цукрового діабету 2 типу на основі поліморфізму FokI

гена VDR, ліпідного профілю та демографічних даних» дослідили етнічні особливості території та визначили параметри, які слід використовувати у моделі. Модель досягла точності від 70 до 80% [6].

«Машина опорного вектора» (SVM) — це контрольований алгоритм машинного навчання, який можна використовувати як для класифікації, так і для задач регресії. Однак він переважно використовується в задачах класифікації. В алгоритмі SVM ми зображуємо кожен елемент даних як точку в n-вимірному просторі (де n — кількість властивих вам властивостей), при цьому значення кожної характеристики є значенням певної координати. Потім ми виконуємо класифікацію, знаходячи гіперплощину, яка дуже добре розрізняє два класи.

У статті від 2019 року «Прогнозування довгострокового діабету 2 типу за допомогою опорного вектора за допомогою перорального тесту на толерантність до глюкози» було отримано результати середню точність 96,80% і чутливість 80,09% [38].

Стаття 2020 року «Діагностичне прогнозування діабету за допомогою векторних опорних машин» показала, що найбільш важливими ознаками є вік, ІМТ та концентрація глюкози в крові. Було впроваджено модель опорних векторів та досягнуто точності 99 відсотків для пацієнтів колумбійського походження та 65,6 відсотків іншого етнічного походження [16].

У випадковому лісі ми створюємо декілька випадкових дерев рішень, які вирішують проблему з різними підмножинами даних. Кожне дерево навчається на випадково обраних функціях та випадково обраних підмножинах даних. Потім, коли потрібно зробити прогноз, кожен випадковий дерево повертає свій прогноз, і ці прогнози об'єднуються в один загальний прогноз шляхом голосування або усереднення.

У статті «Модель прогнозування цукрового діабету 2 типу для пацієнтів з переддіабетом Оману з використанням штучної нейронної мережі та шести класифікаторів машинного навчання» було враховано 11 клінічних ознак. Моделі випадкового лісу та дерева рішень працювали краще, ніж усі інші алгоритми, забезпечуючи точність 98,38% для даних Омана. При використанні

тієї ж моделі та кількості функцій точність, отримана з набором даних Омана, перевищила PID на 9,1%. Аналіз показав, що ефективність діагностики цукрового діабету 2-го типу підвищилась із збільшенням кількості функцій, що допомагає у випадку багатьох відсутніх значень [40].

У роботі «Підхід випадкового лісу для визначення прогнозу ризику та прогностичних факторів діабету 2 типу: дані широкомасштабного медичного обстеження в Японії» досліджено загальну кількість 42 908 осіб, які не отримували лікування діабету з HbA1c <6,5%. Об'єктивною змінною була зміна HbA1c у наступному році. Було використано два аналітичні методи для порівняння прогностичних можливостей: RF як нову модель і багатовимірну логістичну регресію (MLR) як традиційну модель.

RF-модель продемонструвала вищу прогностичну силу для зміни HbA1c, ніж MLR у всіх моделях [44].

Стаття «Прогноз ризику діабету II типу на основі моделі випадкового лісу» продемонструвала використання алгоритму випадкового лісу на даних медичної школи та визначено, що даний метод дає вищу точність ніж інші методи [47].

Random forest є одним з найпопулярніших алгоритмів машинного навчання для задач класифікації, регресії та кластеризації. Він поєднує декілька дерев рішень, які випадковим чином вибираються з декількох підмножин даних.

Алгоритм дерева рішень - це деревоподібна структура, де кожен вузол представляє собою рішення про подальші дії. У бінарних деревах рішень кожен вузол має два нащадки, які представляють собою можливі рішення. Дерева рішень використовуються для вирішення проблем класифікації та регресії. Їх використовують для вирішення різноманітних завдань, таких як визначення рівня захворювання, категоризація вмісту та прийняття рішень.

Random forest є потужним алгоритмом машинного навчання, який має багато переваг. Він працює добре з великими наборами даних, може працювати з даними високої розмірності та показує добрі результати на нових даних.

Таблиця 2.1

Порівняльна характеристика методів машинного навчання*

Метод	Переваги	Недоліки
Логістична регресія	Простота та швидкість роботи, легко інтерпретується	Може бути неефективна для складних взаємозв'язків
Дерево рішень	Добре показує результати на даних з несиметричними класами, можна легко інтерпретувати результати	Схильний до перенавчання та недостатньо універсальний
Нейронні мережі	Висока точність, може працювати зі складними даними	Схильний до перенавчання, складність побудови та інтерпретації
Метод опорних векторів (SVM)	Висока точність, ефективний з обмеженими даними	Складний в побудові та використанні, вимагає налаштування гіперпараметрів
Random Forest	Висока точність, може працювати зі складними даними та великою кількістю змінних, мале схильність до перенавчання	Потребує багато часу для побудови та підготовки даних, складність інтерпретації результатів

*авторська розробка

Random forest є ефективним методом для визначення діабету з кількох причин:

- Він може обробляти як категоріальні, так і числові дані, які часто зустрічаються в дата сетах про діабет.
- Він є потужним алгоритмом, який може працювати з великими обсягами даних, що є важливим у випадку досліджень про діабет.

- Random forest відносно стійкий до шуму та випадкових відхилень у даних, що дозволяє йому показувати гарні результати в реальних умовах.
- Він може автоматично визначати важливість ознак, що допомагає вибрати найбільш значущі ознаки для побудови моделі та покращення її ефективності.
- Random forest може працювати зі складними залежностями між ознаками, що може бути корисним при визначенні складних хвороб, таких як діабет.

Отже, завдяки цим перевагам Random forest може бути ефективним методом для визначення діабету.

2.3. Математичний опис методів машинного навчання

«Машина опорного вектора» (SVM) — це контрольований алгоритм машинного навчання, який можна використовувати як для класифікації, так і для задач регресії. Однак він переважно використовується в задачах класифікації. В алгоритмі SVM ми зображуємо кожен елемент даних як точку в n-вимірному просторі (де n — кількість властивих вам властивостей), при цьому значення кожної характеристики є значенням певної координати. Потім ми виконуємо класифікацію, знаходячи гіперплощину, яка дуже добре розрізняє два класи.

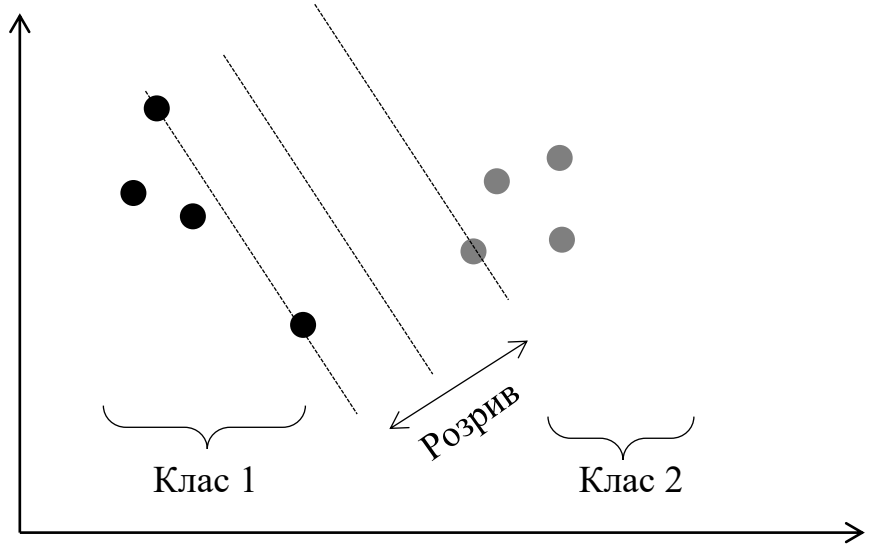

Рис.2.2. SVM метод*

*авторська розробка

Основний принцип роботи машин опорних векторів простий – створити гіперплощину, яка розділяє набір даних на класи. Щоб обрати як саме розташувати лінію, нам необхідно знайти точки, які лежать найближче до обох класів. Ці точки називають опорними векторами. На наступному кроці ми знаходимо близькість між нашою розділовою площиною та опорними векторами. Відстань між точками та лінією поділу відома як розрив. Мета алгоритму SVM — максимізувати саме цей запас. Коли запас досягає свого максимуму, гіперплощина стає оптимальною.

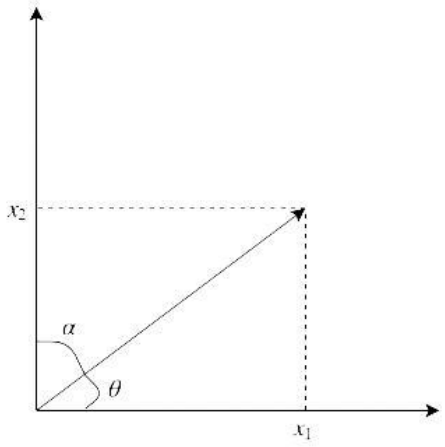

Рис.2.3. Вектор та його складові*

*[62]

Довжина вектора x називається його нормою, яка записується як ||x||. Формула евклідової норми для обчислення норми вектора x = (x1,x2,...,xn):

$$||x|| = \sqrt{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2} \quad (2.1)$$

Напрямок вектора x = (x1,x2) записується як w і визначається як:

$$w = (cos(\theta), cos(\alpha)) \quad (2.2)$$

Добуток двох векторів повертає скаляр.

$$x \times y = ||x||\ ||y||\ cos(\theta) \quad (2.3)$$

Загалом для двох n-вимірних векторів добуток можна обчислити так:

$$x \cdot y = \sum_{i=1}^{n} x_i \times y_u \quad (2.4)$$

Гіперплощина математично визначається за формулою 2.5.

$$w \times x + b = 0 \tag{2.5}$$

$$(ax_1 + x_2 + b = 0) \tag{2.6}$$

Рівняння 2.5 виводиться з двовимірних векторів. Але насправді воно також працює для будь-якої кількості вимірів.

Коли ми маємо гіперплощину, ми можемо використовувати гіперплощину для прогнозування. Визначимо функцію гіпотези h як:

$$H(x_i) = \begin{cases} +1, \ if \ w \times x + b < 0 \\ -1, \ if \ w \times x + b \geq 0 \end{cases} \tag{2.7}$$

Точка вище або на гіперплощині буде класифікуватися як клас +1, а точка під гіперплощиною буде класифікуватися як клас -1.

Таким чином, в основному, мета алгоритму навчання SVM — знайти гіперплощину, яка могла б точно розділити дані. Таких гіперплощин може бути багато. Важливо знайти найкращий варіант, який часто називають оптимальною гіперплощиною.

Логістична регресія - метод класифікації даних у машинному навчанні. Логістична регресія зазвичай використовується там, де є необхідність класифікувати дані на два або більше класів. Існує два види логістичної регерсії: бінарна, багатокласова логістична регресія. Як випливає з назви, двійковий клас має 2 класи: Так/Ні, Правда/Неправда, 0/1 тощо. У мультикласовій класифікації існує більше 2 класів для класифікації даних.

Логістична регресія названа на честь функції, яка використовується в основі методу, логістичної функції. Логістична функція, яку також називають сигмовидною - це S-подібна крива, яка може перетворити будь-яке дійсне число у значення від 0 до 1.

Класифікатор дасть нам набір вихідних даних або класів на основі ймовірності, коли вхідні дані передаються через функцію і повертаються у вигляді оцінки ймовірності від 0 до 1.

Наприклад, є 2 класи, (1 — здорові, 0 —хворі). Обирається пороговое значення в межах від нуля до одиниці, Оберемо умовно поріг 0.5, тоді якщо функція прогнозування повернула значення 0.7, то це спостереження - клас 1

(здорові). Якщо повернене значення 0.2, класифікувати дане спостереження слід як клас 2 (хворі). Формула логістичної регресії представлена нижче (2.8), де S – сигмоїдна функція.

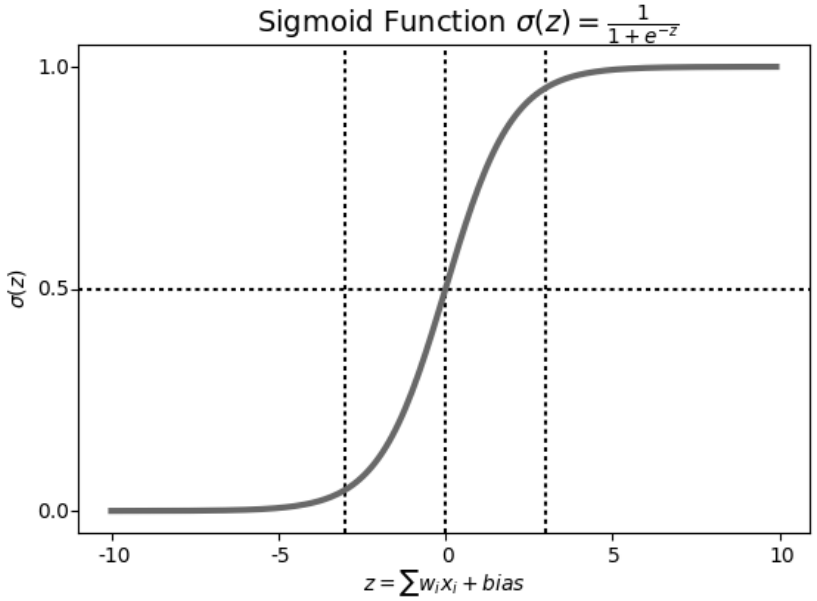

Рис.2.4. Логістична регресія*

*[7]

$$Y = S(a + bX) \tag{2.8}$$

Для зручності візьмемо a + bX = Z, тоді перетворення виглядатимуть наступним чином.

$$S(Z) = \frac{1}{1+e^{-Z}} \tag{2.9}$$

Якщо Z великий і позитивний, S(Z) = 1/(1 + 0) = 1 (приблизно), у разі ж якщо Z великий і негативний, S(Z) = 1/(1 + велика кількість) = 0 (приблизно).

Алгоритм k-найближчих сусідів (KNN) - це простий у реалізації алгоритм машинного навчання з наглядом, який можна використовувати для вирішення проблем як класифікації, так і регресії.

Алгоритм KNN передбачає, що подібні речі існують в безпосередній близькості. Іншими словами, подібні речі знаходяться поруч один з одним.

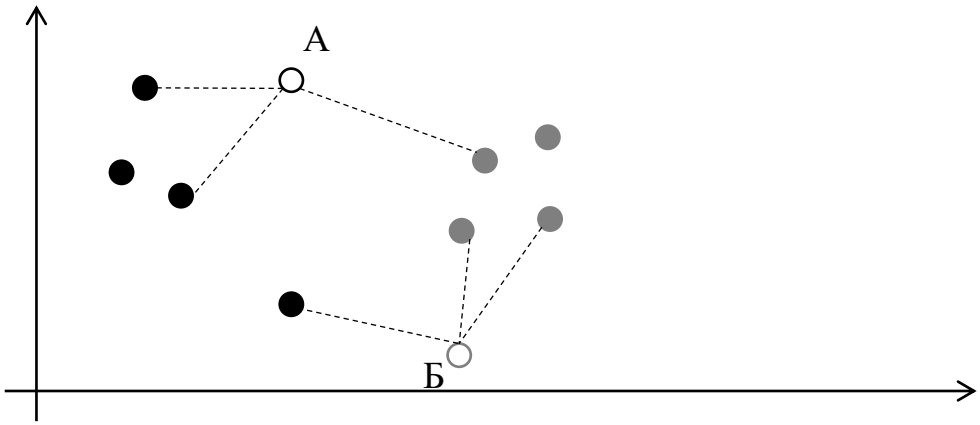

Рис.2.5.Алгоритм KNN*

*авторська розробка

Розглянемо рисунок 2.3., незамальовані точки – це обрані нами елементи, до яких ми шукаємо трьох найближчих сусідів. У точці А два сусіди чорного кольору, тому ми відносимо її також до даного класу. У точці Б – більшість сусідів сірого, тому ми відносимо її до іншого.

Для метрики відстані використовується евклідову метрику.

$$dist((x,y),(a,b)) = \sqrt{(x - a)^2 + (y - b)^2} \qquad (2.10)$$

Евклідову відстань можна просто визначити як найкоротшу між 2 точками незалежно від розмірів. Відповідно до евклідової формули відстані відстань між двома точками на площині з координатами (x, y) і (a, b) визначається як формула 2.10.

Для заданого значення K алгоритм знайде k-найближчих сусідів, а потім призначить клас точці даних, який має найбільшу кількість точок даних з усіх класів K.

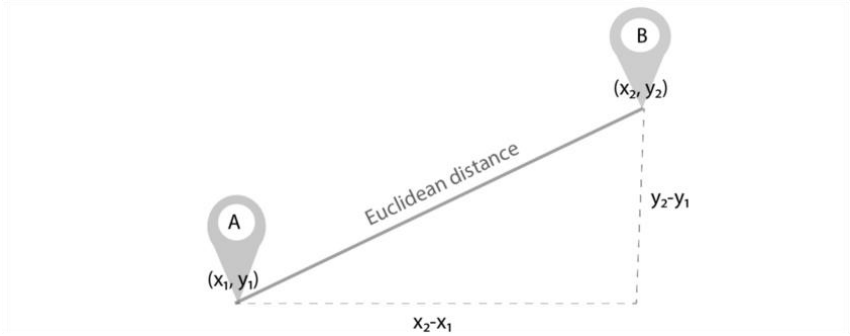

Рис.2.6. Візуалізація евклідової відстані*

*[58]

Після обчислення відстані вхід x отримує клас з найбільшою ймовірністю:

$$P(y = j \mid X = x) = \frac{1}{K}\sum_{i \in A} I\left(y^i = j\right) \quad (2.11)$$

Оскільки у роботі використовується бібліотека Scikit-learn, варто описати який алгоритм використовує вона у методі випадкового лісу.

У документації Scikit-learn зазначено, що він використовує «оптимізовану версію алгоритму CART» [69].

CART розшифровується як дерева класифікації та регресії. Алгоритм створює бінарне дерево — кожен вузол має рівно два вихідних ребра — знаходячи найкращу числову або категоріальну ознаку для розділення за допомогою відповідного критерію домішок. Для класифікації можна використовувати домішки Джині або критерій подвійності. Для регресії CART запровадив зменшення дисперсії за допомогою методу найменших квадратів (середня квадратична помилка).

Оскільки проблема роботи категоріальна, використовується індекс Джинні.

$$\sum_{i=1}^{C} f_i(1 - f_i) \quad (2.12)$$

Де C – це кількість унікальних ознак, f - це частота ознак.

Ще один термін, на який варто звернути увагу, — це «приріст інформації», який використовується з розділенням даних за допомогою ентропії. Він обчислюється як зменшення ентропії після того, як набір даних розділено на атрибут:

Приріст (T, X) = Ентропія (T) — Ентропія (T, X)

T = цільова змінна

X = Функція, на яку потрібно розділити

Ентропія(T,X) = ентропія, обчислена після того, як дані розділені на функцію X

Випадкові ліси (RF) будують багато окремих дерев рішень під час навчання. Прогнози з усіх дерев об’єднуються, щоб зробити остаточний прогноз.

Важливість функції обчислюється як зменшення домішки вузла, зважене на ймовірність досягнення цього вузла. Імовірність вузла можна розрахувати як кількість зразків, які досягають вузла, поділену на загальну кількість зразків. Чим вище значення, тим важливіша функція.

$$f_{i_i} = \frac{\sum_j node\ j\ splits\ on\ feature\ i\ n_{i_j}}{\sum_{k \in all\ nodes\ n_{i_k}}} \tag{2.13}$$

2.4. Висновки до другого розділу

У другому розділі було проаналізовано різні алгоритми машинного навчання, їх класифікацію та особливості застосування. Було надано математичний опис методів машинного навчання, а саме логістична регресія, SVM та KNN методи, та метод випадково лісу. Проведено комплексний аналіз переваг та недоліків кожного з методів. Опрацьовано актуальний стан методів імплементації методів.

Також було надано формули та приклади використання кожного з цих методів. Наприклад, була надана формула для логістичної регресії та описано її застосування для бінарної класифікації.

Після того, як було проаналізовано різні ММН, було проведено дослідження статей, в яких використовувалися ці методи. Зокрема, було проаналізовано статті, що стосуються визначення ДДТ.

В результаті аналізу було обрано метод випадкового лісу (Random Forest) як найбільш оптимальний метод для вирішення задачі визначення ДДТ. Це пов'язано з тим, що метод Random Forest відносно простий у реалізації, дозволяє враховувати велику кількість ознак та має досить високу точність прогнозування.

РОЗДІЛ 3. ПОБУДОВА МОДЕЛІ ПРОГНОЗУВАННЯ ДДТ ЗА ДОПМОГОЮ МЕТОДА ВИПАДКОВОГО ЛІСУ

3.1. Формалізація бази знань захворювання ДДТ

Формалізація бази знань захворювання ДДТ є важливим кроком у покращенні діагностики та лікування цієї хвороби. Цей процес включає в себе створення бази даних, яка містить інформацію про причини, симптоми, діагностику та лікування цього захворювання.

Формалізація бази знань дозволяє ефективніше використовувати отриману інформацію та забезпечити більш точну та швидку діагностику. Крім того, вона сприяє покращенню зберігання та обробки медичної інформації, зниженню кількості помилок в діагностиці та лікуванні та, як наслідок, покращенню якості життя пацієнтів з діабетом другого типу.

Запропонована база знань включає два набори даних: числовий та категоріальний.

В ході дослідження було вирішено використати дані з репозиторію для тренування та тестування алгоритму визначення діабету [71].

Цей датасет має такі переваги перед іншими:

Реалістична інформація: датасет містить дані про ризик розвитку діабету на ранній стадії, що є реальними медичними даними, а не вигаданими.

Велика кількість змінних: датасет містить велику кількість змінних, які можна використовувати для аналізу ризиків та іншої статистичної обробки даних.

Легко доступний: цей датасет доступний для безкоштовного завантаження та використання в дослідженнях.

Перевірений: датасет використовувався в декількох наукових дослідженнях та вже пройшов перевірку на достовірність та точність.

Потенційна прогностична цінність: аналіз даних з цього датасету може допомогти визначити ризики розвитку діабету на ранній стадії та вжити

профілактичні заходи для зменшення цих ризиків.

Різноманітність даних: датасет містить інформацію про пацієнтів різних вікових груп, статей, етнічних груп, що дає можливість провести детальний аналіз взаємозв'язку між цими факторами та ризиком розвитку діабету.

Бінарні дані, які легко зібрати у пацієнта, на противагу кількісним аналізам.

Робота над даними включає в себе створення code book, попередню обробку даних, очищення та кодування категоріальних змінних та видалення відсутніх значень, вибір найбільш інформативних змінних, що допоможуть покращити якість моделі та знизити її складність. Для відбору можуть бути використані різноманітні методи, такі як важливість ознак (feature importance), метод головних компонент (PCA) та аналіз взаємозв'язку ознак (feature correlation). Надалі дані розбиваються на тренувальні та тестові, використовуючи випадкове розбиття чи кросс-валідацію. Надалі варто обрати один або декілька алгоритмів, натренувати та перевірити отриману модель. У додатку А наведена блок схема процесу створення технології визначення діабету методом випадкового лісу. Основні етапи алгоритму - збір, очищення та підготовка даних, вибір ознак, відбір ознак з побудовою моделі, оцінка точності моделі та її покращення з використанням PCA. Результати оцінки точності порівнюються, інтерпретуються та розробляється рекомендації. Кінцевий продукт - система для діагностики ДДТ. У додатку Б наведений Code Book для використаного набору даних

У роботі використані Jupiter Notebook, python та відповідні бібліотеки.

В першу чергу було завантажено набір даних та проведено попередню обробку, дані є доволі чистими, відсутні пропущені значення. Було проведено дослідницький аналіз данних, визначено розподіли змінних.

Цей датасет містить інформацію про 520 пацієнтів, що мають або не мають цукрового діабету. Загалом, 320 пацієнтів у датасеті мають діабет, а 200 пацієнтів не мають.

Стать пацієнтів також представлена в цьому датасеті. Загалом, 328

пацієнтів у датасеті є чоловіками, а 192 - жінками.

Датасет також містить інформацію про вік пацієнтів. Вік пацієнтів варіюється від 16 до 90 років, а середній вік пацієнтів у датасеті становить близько 48 років.

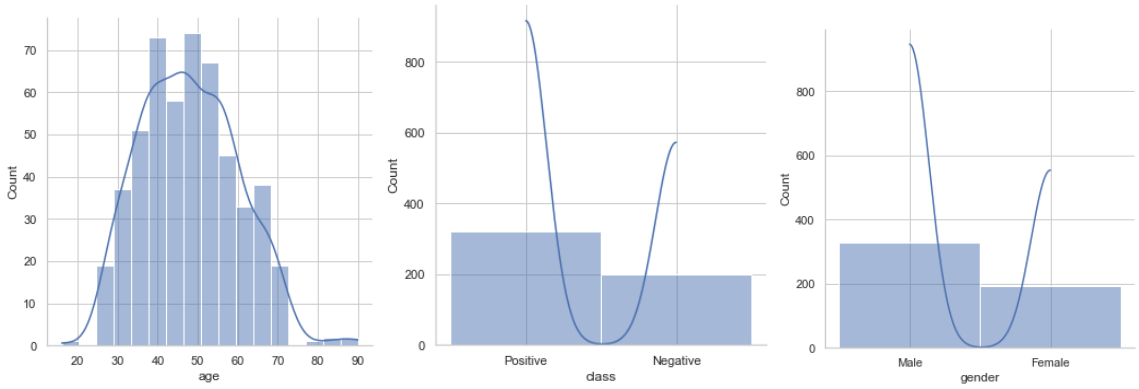

Рис. 3.1 Дескриптивний аналіз набору даних*

*авторська розробка

Для роботи з бінарними даними надалі було проведено маппінг бінарних значень на числові. Надалі було проведено корреляційний аналіз та визначено найбільш впливові ознаки на змінну класс.

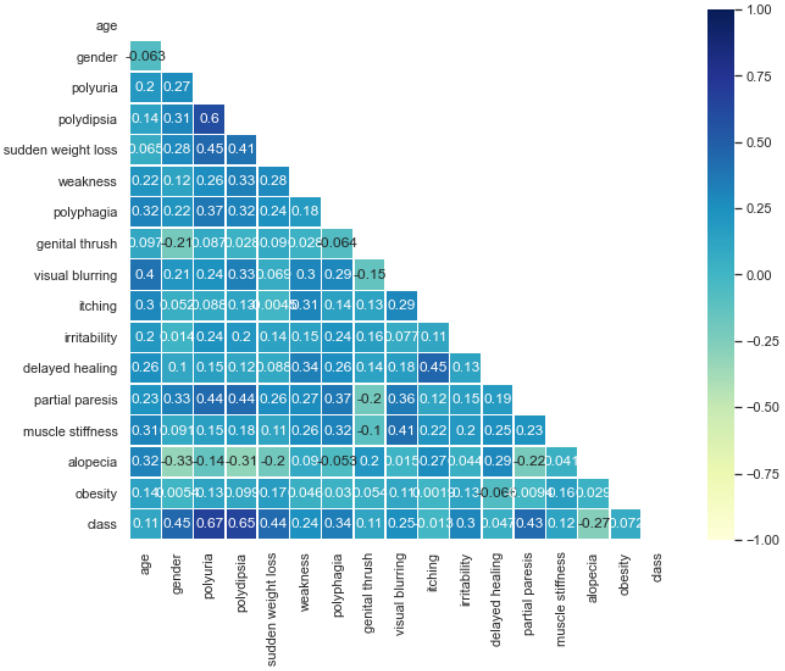

Рис.3.2 Кореляційний аналіз ознак набору даних з результуючою ознакою клас*

*авторська розробка

Було вирішено відхили ті ознаки, які в абсолютному значенні були менші за 0.3.

Таблиця 3.1

Кореляція з результуючою ознакою*

Номер	Ознака	Кореляція
1	class	1
2	polyuria	0.665922
3	polydipsia	0.648734
4	gender	0.449233
5	sudden weight loss	0.436568
6	partial paresis	0.432288
7	polyphagia	0.342504
8	irritability	0.299467
9	alopecia	0.267512
10	visual blurring	0.2513
11	weakness	0.243275
12	muscle stiffness	0.122474
13	genital thrush	0.110288
14	age	0.108679
15	obesity	0.072173
16	delayed healing	0.04698
17	itching	0.013384

*авторська розробка

Надалі було сформовано новий графік кореляції, з якого видно найбільш вагомі ознаки для визначення діабету.

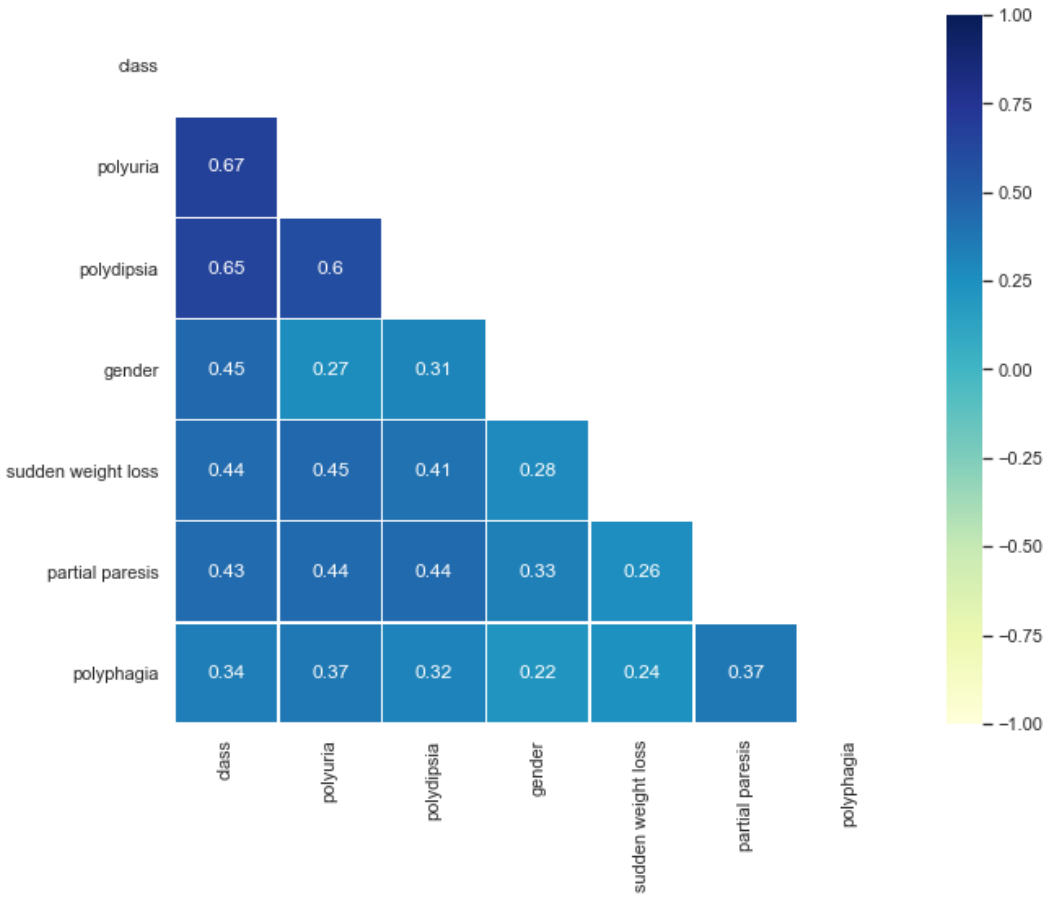

Рис.3.3 Відібрані ознаки та їх кореляція з результуючою ознакою клас*

*авторська розробка.

Свіжі дані у області діабету було взято з набору даних від іракського суспільства про 1000 пацієнтів (з міської медичної лікарні). Дані включають: стать, вік, сечовину, Cr (співвідношення креатиніну), HbA1c (глікогемоглобін), холестерин, TG (тригліцериди), HDL (ліпопротеїни високої щільності), LDL (ліпопротеїни низької щільності), VLDL (дуже низький рівень щільності), IMT (індекс маси тіла), клас захворювання діабету пацієнта (може бути діабетичним, недіабетичним або прогнозованим діабетом).

Варто відзначити, що дані текстового формату були замінені на числову репрезентацію: так жінки у датасеті відображені як 1, а чоловіки як 0 відповідно. Клас людей з діабетом рівний 1, без діабету 0, а переддіабетичний стан був видалений.

У наборі даних в нас представлено 1000 записів, водночас пізніше очікується зменшення релевантних даних, адже ще не проведена попередня

обробка даних.

Таблиця 3.2

Опис набору даних*

	count	mean	std	min	25%	50%	75%	max
age	947	54.101373	8.499612	20	51	55	59	79
urea	947	5.159074	2.975024	0.5	3.7	4.6	5.7	38.9
cr	947	69.103485	60.862961	6	48	60	73	800
hba1c	947	8.408617	2.54404	0.9	6.8	8.1	10.2	16
chol	947	4.878691	1.313356	0	4	4.8	5.6	10.3
tg	947	2.362101	1.417275	0.3	1.5	2	2.9	13.8
hdl	947	1.209081	0.672423	0.2	0.9	1.1	1.3	9.9
ldl	947	2.616304	1.127316	0.3	1.8	2.5	3.3	9.9
vldl	947	1.903485	3.757012	0.1	0.7	1	1.5	35
bmi	947	29.893897	4.869852	19	27	30	33	47.75
class	947	0.891235	0.311508	0	1	1	1	1

*авторська розробка

Набір даних містить 947 зразків з наступними ознаками:

Вік (age): Середнє значення - 54.10 - це середній вік пацієнтів у досліджуваній популяції. Мінімальне значення - 20 - наймолодший вік серед пацієнтів, а максимальне значення - 79 - найстарший вік серед пацієнтів.

Урея (urea): Середнє значення - 5.16 - це середній рівень уреї в крові пацієнтів. Мінімальне значення - 0.5 - найнижчий рівень уреї, що був виміряний, а максимальне значення - 38.9 - найвищий рівень уреї серед пацієнтів. Урея - це продукт обміну амінокислот, який виробляється в печінці та використовується для оцінки функції нирок та стану обміну білка в організмі.

Креатинін (creatinine): Середнє значення - 69.10 - це середній рівень креатиніну в крові пацієнтів. Мінімальне значення - 6 - найнижчий рівень креатиніну, що був виміряний, а максимальне значення - 800 - найвищий рівень

креатиніну серед пацієнтів. Креатинін - це продукт метаболізму креатину, який виробляється в м'язовій тканині та використовується для оцінки функції нирок та стану м'язової маси в організмі.

Глюкоза (glucose): Середнє значення - 103.40 - це середній рівень глюкози в крові пацієнтів. Мінімальне значення - 70 - найнижчий рівень глюкози, що був виміряний, а максимальне значення - 260 - найвищий рівень глюкози серед пацієнтів. Глюкоза - це основний джерело енергії для організму, а рівень глюкози в крові може бути використаний для оцінки рівня цукрового діабету та функції обміну вуглеводів.

Холестерин (cholesterol): Середнє значення - 5.16 - це середній рівень загального холестерину в крові пацієнтів. Мінімальне значення - 2.1 - найнижчий рівень холестерину, що був виміряний, а максимальне значення - 8.0 - найвищий рівень холестерину серед пацієнтів. Холестерин - це тип ліпідів, який виконує ряд важливих функцій в організмі, а його рівень в крові може бути використаний для оцінки ризику розвитку серцево-судинних захворювань.

Тригліцериди (triglycerides): Середнє значення - 1.48 - це середній рівень тригліцеридів в крові пацієнтів. Мінімальне значення - 0.5 - найнижчий рівень тригліцеридів, що був виміряний, а максимальне значення - 4.9 - найвищий рівень тригліцеридів серед пацієнтів. Тригліцериди - це форма жирів, які зберігаються в організмі та можуть бути використані як джерело енергії. Високі рівні тригліцеридів можуть бути пов'язані з ризиком розвитку серецевих захворювань.

АЛТ (alanine aminotransferase): Середнє значення - 25.60 - це середній рівень АЛТ в крові пацієнтів. Мінімальне значення - 10 - найнижчий рівень АЛТ, що був виміряний, а максимальне значення - 60 - найвищий рівень АЛТ серед пацієнтів. АЛТ - це фермент, який зазвичай знаходиться в клітинах печінки, а величина його рівня в крові може вказувати на функціональний стан печінки та можливі печінкові захворювання.

Після перевірки на кореляцію між ознаками та результуючою виявилось, що є логічним залишити лише три ознаки:

Зважений коефіцієнт маси тіла (BMI): Це ознака, що визначається на основі ваги та зросту пацієнта і використовується для оцінки відносного розподілу маси тіла. Він обчислюється за формулою BMI = вага (кг) / (зріст (м) * зріст (м)).

Середнє значення BMI може варіюватися в залежності від регіональних стандартів, але зазвичай вважається, що нормальний діапазон BMI знаходиться в межах 18,5-24,9.

Гемоглобін A1C (HbA1c): Це вимірювання рівня гемоглобіну A1C в крові, що використовується для оцінки рівня глікемії (рівня цукру в крові) протягом останніх 2-3 місяців. Вищі значення HbA1c можуть вказувати на погане контролювання рівня цукру в крові, що може бути пов'язано з діабетом типу 1 або 2, а також збільшенням ризику розвитку ускладнень діабету, таких як пошкодження нирок, нервів, очей та серця.

Вік (age): з віком зменшується резервні можливості організму для регенерації та пристосування до стресів.

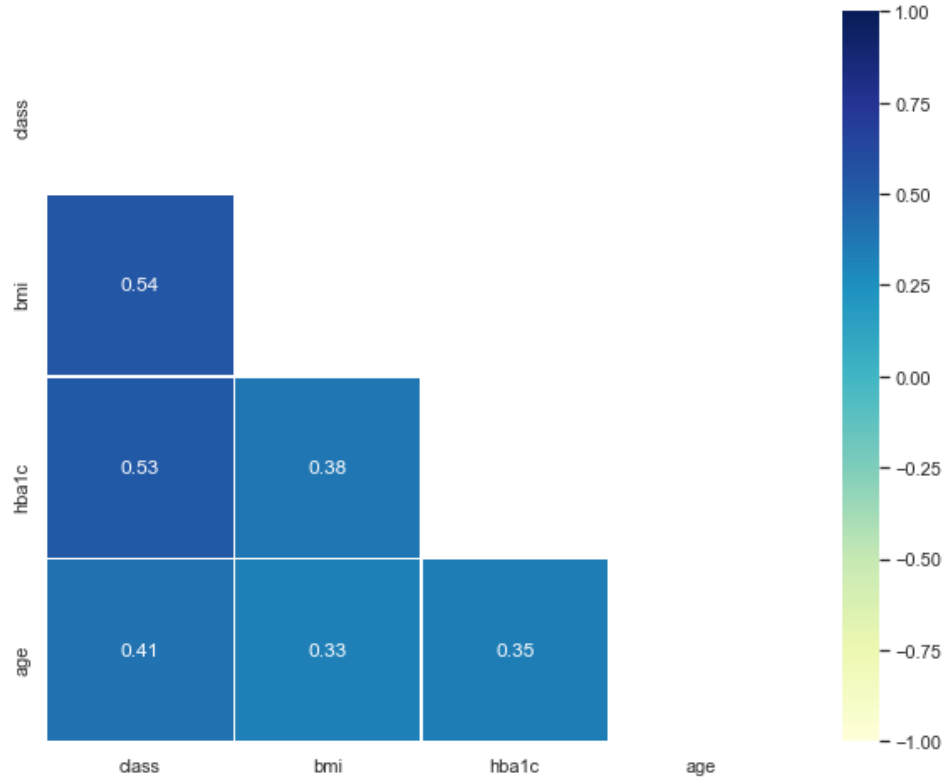

Рис.3.4. Матриця кореляції*

*авторська розробка

Таблиця 3.3

Кореляція ознак із результуючою ознакою*

	Feature	Correlation
0	class	1.000000
1	bmi	0.539671
2	hba1c	0.528732
3	age	0.405810
4	tg	0.181722
5	chol	0.162275
6	vldl	0.089292
7	urea	0.056216
8	cr	0.036222
9	hdl	0.009410
10	ldl	0.002771

*авторська розробка

Продукційна модель представлення знань є однією з найпоширеніших і в даному випадку відповідає проблематиці експертної системи. Представлення знань за допомогою правил-продукцій має в деяких відносинах подібність із правилами виводу логічних моделей. Це дозволяє за допомогою продукцій виконувати ефективний вивід і, крім того, завдяки природній аналогії процесу міркувань людини дані моделі наочніше відбивають знання.

Широке застосування продукційних моделей визначається наступними основними перевагами:

- універсальністю, практично будь-яка область знань може бути представлена в продукційній формі;
- модульностью, кожна продукція являє собою елемент знань про предметну область, видалення одних і додавання інших продукцій
- виконується незалежно;

- декларативністю, продукції визначають ситуації, предметної області, а не механізму керування;
- природністю процесу виводу, що багато в чому аналогічний процесу міркувань експерта;
- асинхроністю і природним паралелізмом

Однак продукційні моделі не свободні від недоліків: – процес виводу має низьку ефективність, тому що при великому числі продукцій значна частина часу затрачається на невиробничу перевірку умов застосування правил; – перевірка несуперечності системи продукцій стає дуже складною через недетермінованності вибору виконуваної продукції з конфліктної множини.

База знань у продукційній моделі — це сукупність бази фактів і бази правил. Кожне продукційне правило в БЗ втілює автономну частину експертних знань одержаних від експерта при набутті знань вручну або використовуючи методи автоматичного видобування знань.

Окреме продукційне правило може бути розроблене та модифіковане незалежно від інших правил. Правила можна розглядати, в певному сенсі, як симуляцію когнітивної поведінки експерта в певній проблемній області.

Згідно з цим поглядом, правила є не лише чистим формалізмом для представлення знань в комп'ютері; швидше вони представляють модель фактичної людської поведінки.

Продукційна модель найчастіше використовується в промислових експертних системах.

Наприклад, у медичній експертній системі правила «якщо, то» можуть використовуватися для встановлення взаємозв'язків між симптомами і діагнозами. Під час виведення реальний симптом зіставляється з тим, які є в лівих частинах правил і в разі збігу права частина відповідного правила вважається можливим діагнозом.

Модель випадкового лісу має багато дерев, тому запропоновано розглянути одне із дерев для категоріальних даних оброблених методом головних компонент.

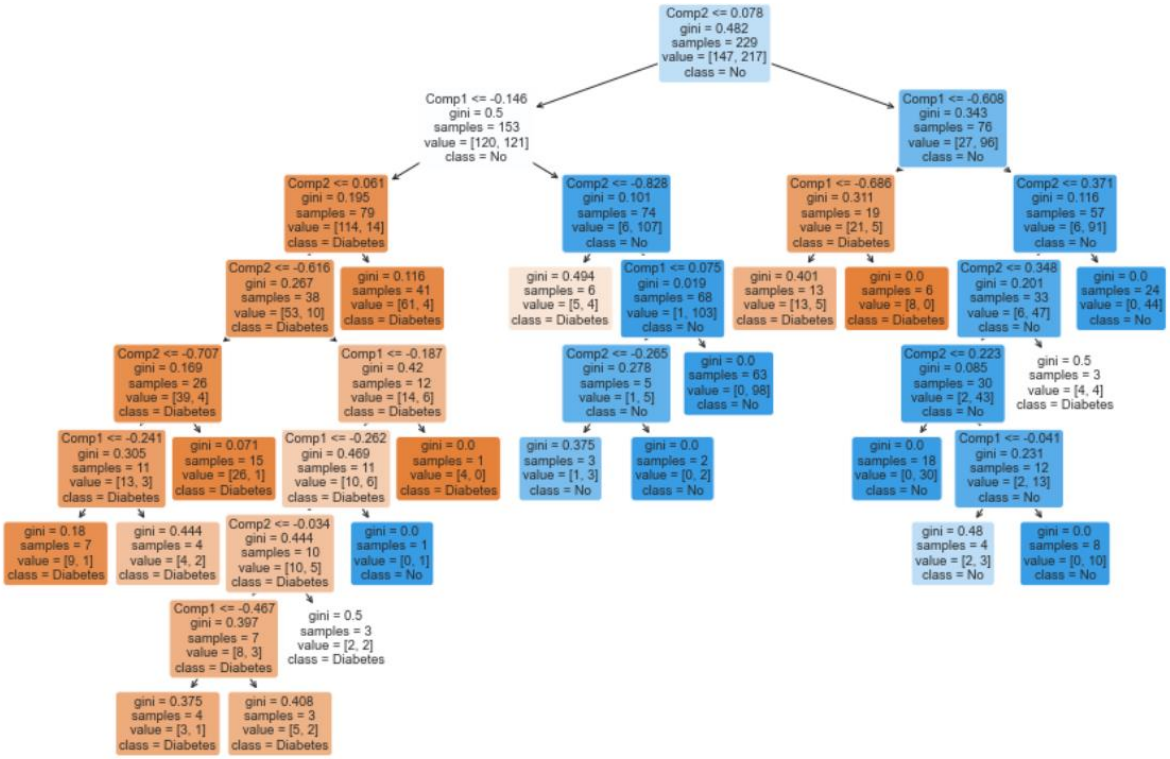

Рис.3.5. Приклад дерева*

*авторська розробка

Наведемо також приклад дерева без використання компонент.

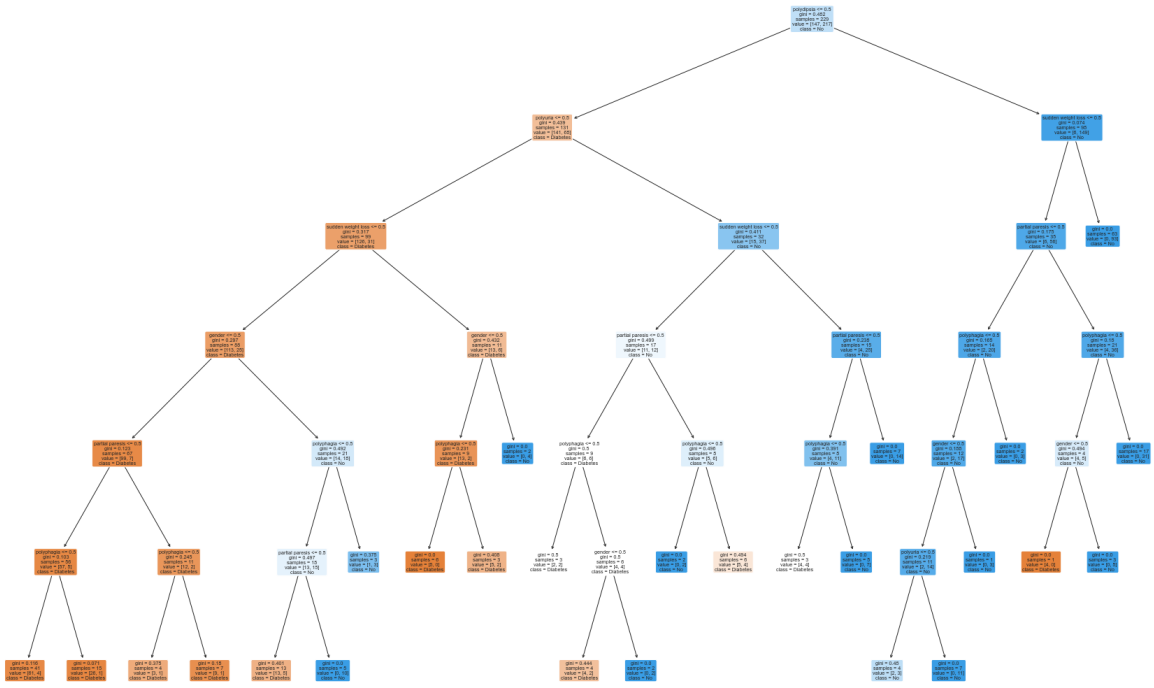

Рис.3.6. Приклад дерева для ознак*

*авторська розробка

3.2. Вибір засобів для реалізації технології визначення ДДТ

Розглядаючи мови програмування, що використовуються у машинному навчанні, можемо відзначити, що Python лідує: 57% науковців даних і розробників машинного навчання використовують його. Python часто порівнюють з R, але за популярністю R займає четверте місце за загальним використанням (31%) і п’яте за визначенням пріоритетів (5%).

Насправді R є мовою з найнижчим співвідношенням пріоритетності та використання серед п’яти, і лише 17% розробників, які її використовують, дають їй пріоритет. Це означає, що в більшості випадків R є додатковою мовою, а не першим вибором. Такий самий коефіцієнт для Python становить 58%, що є найвищим серед п'яти мов.

Python є не тільки найбільш широко використовуваною мовою, але й основний вибір для більшості користувачів. C/C++ є другим після Python, як у використанні (44%), так і в пріоритеті (19%). Java слідує за C/C++ дуже близько, тоді як JavaScript займає п’яте місце у використанні, хоча з дещо кращою продуктивністю визначення пріоритетів, ніж R (7%). Також використовуються в машинному навчанні мови Julia, Scala, Ruby, Octave, MATLAB і SAS [64].

Бібліотеки Python — це колекції модулів, які містять корисні коди та функції, що усуває необхідність писати їх з нуля. Існують десятки тисяч бібліотек Python, які допомагають розробникам машинного навчання, а також професіоналам, які працюють у галузі даних, візуалізації даних тощо.

Python є найкращою мовою для машинного навчання, оскільки її синтаксис і команди тісно пов’язані з англійською, що робить її ефективною та легкою для вивчення. У порівнянні з C++, R, Ruby та Java, Python залишається однією з найпростіших мов, що забезпечує доступність, універсальність і портативність. Він може працювати практично на будь-якій операційній системі чи платформі.

Дослідження проблематики роботи здійснюватиметься мовою програмування Python. Найбільшою перевагою даної мови є її спільнота та кількість підходящих бібліотек.

Python є однією з найбільш часто використовуваних для створення програмного забезпечення на основі ML.

Він має швидку криву навчання, забезпечує абсолютно безболісну взаємодію з різними системами керування базами даних і легко інтегрується з різними програмними інструментами, що спеціалізуються виключно на створенні алгоритмів машинного навчання.

Деякі з бібліотек, які підтримуються в Python для машинного навчання: Tensorflow для глибокого навчання, Numpy для математичних операцій, Pandas для файлових операцій, Pytorch для пакета глибокого навчання, Sklearn для алгоритмів класифікації та регресії, OpenCV і Dlib для комп'ютерного зору і Matplotlib, Seaborn для візуалізації даних.

З усіма цими перевагами, Python також має кілька недоліків: він відносно повільніший, ніж інші мови, такі як C++, а також важко підтримувати багатопотоковість [10].

Розглянемо бібліотеки дещо детальніше:

NumPy - добре відомий пакет для обробки масивів загального призначення. Велика колекція математичних функцій високої складності робить NumPy потужним для обробки великих багатовимірних масивів і матриць. NumPy дуже корисний для роботи з лінійною алгеброю, перетвореннями Фур’є та випадковими числами. Інші бібліотеки дуже часто використовують дану бібліотеку як основу.

SciPy - пропонує модулі для лінійної алгебри, оптимізації зображень, інтерполяції інтеграції, спеціальних функцій, швидкого перетворення Фур'є, обробки сигналів і зображень, розв'язування звичайних диференціальних рівнянь та інших обчислювальних завдань у науці та аналітиці. Базовою структурою даних, що використовується SciPy, є багатовимірний масив, наданий модулем NumPy. SciPy залежить від NumPy для підпрограм маніпуляції з масивом. Бібліотека SciPy була створена для роботи з масивами NumPy разом із наданням зручних та ефективних числових функцій.

Scikit-learn - має широкий спектр контрольованих і неконтрольованих

алгоритмів навчання, які працюють на узгодженому інтерфейсі на Python. Бібліотеку також можна використовувати для штучного інтелекту та аналізу даних. Основні функції машинного навчання, з якими може працювати бібліотека Scikit-learn, — це класифікація, регресія, кластеризація, зменшення розмірності, вибір моделі та попередня обробка.

PyTorch - має ряд інструментів і бібліотек, які підтримують комп'ютерний зір, машинне навчання та обробку природної мови. Бібліотека PyTorch є відкритим вихідним кодом і заснована на бібліотеці Torch. Найважливішою перевагою бібліотеки PyTorch є її простота навчання та використання.

Pandas - стала найпопулярнішою бібліотекою Python, яка використовується для аналізу даних з підтримкою швидких, гнучких і виразних структур даних, призначених для роботи як з «реляційними», так і з простими даними. Pandas сьогодні — це неминуча бібліотека для вирішення практичного аналізу даних у реальному світі на Python.

Pandas дуже стабільний і забезпечує оптимізовану продуктивність. Бекенд-код написаний виключно на C або Python.

Таблиця 3.4

Порівняльна таблиця методів машинного навчання на обраному сеті даних*

Method	**precision**	**recall**	**f1-score**
SVM	0,81	0,85	0,83
LR	0,67	0,65	0,65
KNN	0,68	0,72	0,70
RF	0,98	0,96	0,95

*авторська розробка

Крім того, Scikit-learn має широку спільноту користувачів, яка підтримує та розвиває бібліотеку, що забезпечує надійність та стабільність її роботи. Тому, використання Scikit-learn може значно спростити та прискорити розробку

системи визначення діабету методами машинного навчання.

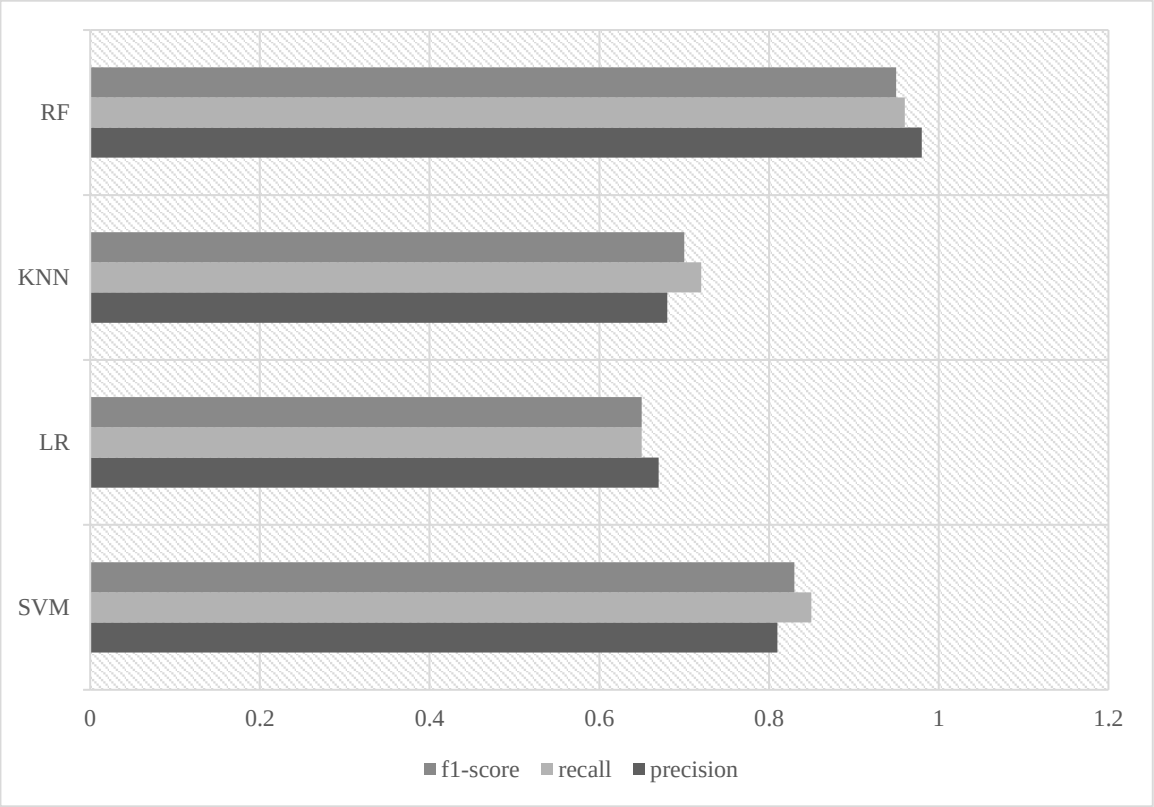

Рис.3.6.Графічне порівняння метрик методів*

*авторська розробка

Бібліотека Scikit-learn є однією з найбільш популярних бібліотек для машинного навчання в середовищі Python. Вона містить багато інструментів для класифікації, регресії, кластеризації та інші ММН, що дозволяє легко тестувати різні алгоритми та порівнювати їх ефективність.

Scikit-learn також надає зручний інтерфейс для обробки та підготовки даних для машинного навчання, таких як зменшення розмірності даних, кодування категоріальних змінних, обробка пропущених значень та інше.

Крім того, Scikit-learn має широку спільноту користувачів, яка підтримує та розвиває бібліотеку, що забезпечує надійність та стабільність її роботи. Тому, використання Scikit-learn може значно спростити та прискорити розробку системи визначення діабету методами машинного навчання.

Було протестовано досліджені алгоритми на наборі даних іракського суспільства та визначено, що метод випадкового лісу виявляє себе як найточніший. Саме тому подальша розробка була запроваджена даним методом.

3.3. Побудова моделі машинного навчання методом випадкового лісу для визначення ДДТ

Було побудовано алгоритм машинного навчання використовую метод випадкового лісу з такими параметрами як random_state початкове число для генератора випадкових чисел, який використовується для поділу набору даних на набори для навчання та тестування. Встановлення значення 42 є лише умовністю. Вибір початкового номера не має значення, якщо він однаковий у різних циклах алгоритму. Це допомагає забезпечити відтворюваність результатів.

Параметр test_size визначає частку набору даних, яка використовується для тестування моделі, а решта даних використовується для навчання. Test_size 0,3 означає, що 30% даних використовується для тестування та 70% для навчання. Це загальне значення, яке використовується в експериментах з машинним навчанням. Для аналізу результатів побудованих моделей використаємо такі методи як confusion matrix та її розширення.

У ході побудови моделі було використано ряд функцій. Опишемо їх роботу.

В ході виокремлення важливих ознак класу було використано метод .corr() в Python для обчислення кореляції між двома змінними. Він приймає на вхід дві змінні, які потрібно порівняти, і повертає значення кореляції між ними. Кореляція - це статистичний параметр, який вказує на ступінь залежності між двома змінними.

Метод .corr() може бути використаний з об'єктом типу DataFrame бібліотеки Pandas для порівняння кількох змінних між собою. Результатом його виконання є таблиця кореляційних коефіцієнтів між всіма парами змінних. Зазвичай кореляційні коефіцієнти представлені в діапазоні від -1 до 1. Значення -1 вказує на повну обернену залежність між змінними, 0 - на відсутність залежності, а 1 - на повну пряму залежність між змінними.

Метод .corr() в Python використовує кореляційний коефіцієнт Пірсона за замовчуванням, який вимірює лінійну залежність між двома змінними. Однак, цей метод дозволяє також використовувати інші кореляційні коефіцієнти, такі як Спірмена, Кендалла тощо.

Надалі використовуємо бібліотеку Scikit-learn для створення класифікатора Random Forest, який буде використовуватись для передбачення класу на основі набору ознак.

Спочатку набір даних розділяється на тренувальний і тестовий набори за допомогою методу train_test_split(), де test_size=0.3 означає, що 30% даних будуть використовуватись для тестування.

Функція train_test_split() з модуля sklearn.model_selection використовується для розділення датасету на навчальні та тестові дані. Вона приймає вхідні дані та розмір тестового набору даних і повертає набір вихідних даних (X_train, X_test, y_train, y_test).

X_train та y_train - це набір навчальних даних, які будуть використовуватися для навчання моделі, X_test та y_test - це набір тестових даних, які будуть використовуватися для перевірки точності навчання моделі.

Далі створюється об'єкт класифікатора Random Forest з параметрами n_estimators=100 і random_state=42.

У параметра n_estimators в алгоритмі RandomForestClassifier вказується, скільки дерев рішень повинно бути побудовано в лісі. Чим більше дерев, тим складніша модель і тим більше часу займає її побудова та прогнозування. Однак, більш складні моделі можуть мати кращу точність на тестових даних.

Зазвичай, для встановлення оптимального значення параметра n_estimators рекомендується проводити експерименти та порівнювати результати виконання моделі для різних значень. Проте, значення 100 часто використовується як типове для даного параметра, оскільки воно забезпечує добру точність прогнозування без великих затрат на обчислення.

Після цього класифікатор навчається на тренувальному наборі за допомогою методу fit().

Нарешті, з класифікатором виконуються передбачення на тестовому наборі за допомогою методу predict() і результати зберігаються у змінній y_pred. Крім того, також імпортуються бібліотеки для обчислення метрик, таких як accuracy_score та confusion_matrix, які надалі використані для оцінки результатів класифікації.

Матриця плутанини — це матриця, яка дозволяє візуалізувати продуктивність класифікаційних моделей машинного навчання. За допомогою цієї візуалізації можна краще зрозуміти, як працює модель машинного навчання. Метою створення та побудови матриці плутанини є перевірка точності моделі машинного навчання. Буде добре візуалізувати точність у відсотках, а не використовувати лише число.

Розширені показники матриці плутанини, це точність (або влучність), повнота, f1 значення. Точність відповідає на запитання: «Коли прогноз має даний результат, як часто він правильний?». Тобто влучність є числом правильних результатів, поділеним на число всіх повернених результатів. Повнота ж є числом правильних результатів, поділеним на число результатів, які мало би бути повернуто. Міра, яка поєднує влучність та повноту, є середнє гармонійне влучності та повноти, традиційна F-міра.

	precision	recall	f1-score	support
0	0.85	0.96	0.90	54
1	0.98	0.91	0.94	102
accuracy			0.93	156
macro avg	0.92	0.94	0.92	156
weighted avg	0.94	0.93	0.93	156

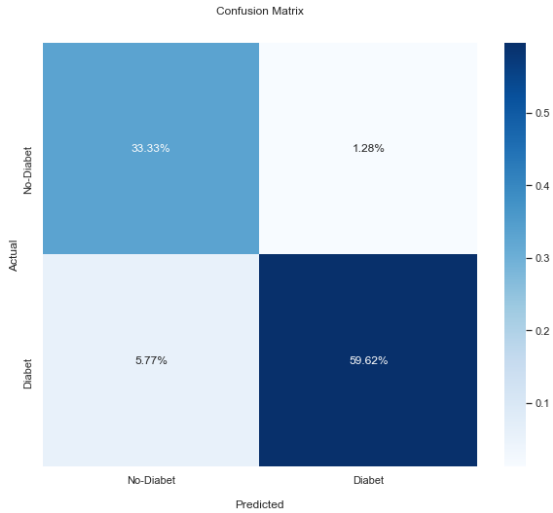

Рис. 3.7. Результати створеної моделі*

*авторська розробка

Хоча модель має високу точність, було вирішено спробувати підвищити її за допомогою методу головних компонент.

Метод головних компонент (Principal Component Analysis або PCA) - це метод зменшення розмірності даних. Головна ідея полягає в тому, щоб знайти нову систему координат для даних, в якій вони будуть більш інформативними. Це здійснюється шляхом перетворення початкових ознак у нові ознаки, які називаються головними компонентами.

Головні компоненти визначаються таким чином, щоб перша компонента захоплювала якнайбільше можливої дисперсії (варіації) даних, друга компонента захоплювала якнайбільше дисперсії, яку не захопила перша компонента, і так далі. В результаті, головні компоненти упорядковані за спаданням їх варіації, тобто перша головна компонента має найбільшу варіацію, друга - меншу, і так далі.

Основна перевага методу головних компонент полягає в тому, що він дозволяє зменшити кількість ознак, не втрачаючи при цьому значної кількості інформації, а також візуалізувати дані у просторі з меншою кількістю вимірів. Це зменшує розмірність простору даних, зберігаючи при цьому їх головні характеристики. Застосування методу головних компонент може покращити якість моделей машинного навчання і зменшити вплив шуму на дані.

Використовуючи даний метод ми змогли підвищити точність наданих прогнозів алгоритму до 93,5%.

Спробуємо відтестувати створений алгоритм визначення ДДТ на умовному кейсі взаємодії лікаря та пацієнта.

Умовно до лікаря звернувся пацієнт з певним анамнезом і лікар має уточнити такі питання:

- Polyuria: Чи спостерігає пацієнт збільшення кількості виділеної сечі.
- Polydipsia: Чи відчуває пацієнт надмірну спрагу, тобто постійне інтенсивне бажання пити.
- Gender: Стать людини, тобто жіноча або чоловіча.

- Sudden weight loss: Різке зменшення ваги, тобто втрата ваги без очевидних причин.
- Partial paresis: Часткова пареза, тобто втрата часткової здатності контролювати рухи.
- Polyphagia: Надмірний апетит, тобто постійне інтенсивне бажання їсти.

Отримані відповіді варто внести до алгоритму. Давайте припустимо, що наш пацієнт - це жінка, яка має такі ознаки:

- polyuria (часте сечовиділення): так
- polydipsia (постійна спрага): так
- gender (стать): жінка
- sudden weight loss (різке схуднення): так
- partial paresis (часткова пареза): так
- polyphagia (постійний голод): так

Було створено нові дані, оброблено методом головних компонент і завантажено до моделі, яка видала що жінка має діабет з вірогідністю 95%.

Алгоритм, хоча і є точним і влучним, але працює лише з суб’єктивними відчуттями людини, або оглядом медичного працівника. Для навчання моделі на цифрових значеннях використаємо той самий алгоритм, проте на іншому наборі даних.

Результуюча модель має результат точності 97.2%, із застосування методом головних компонент 94%, такий результат був очікуваний, адже ознак в цілому не велика кількість тому звужуючи їх модель втратила точність.

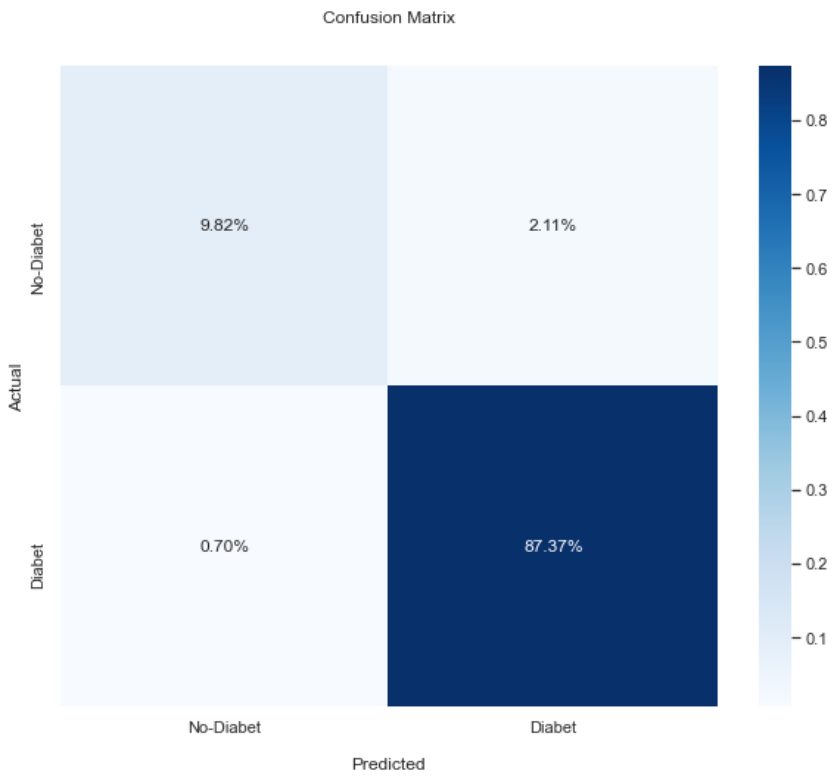

Рис.3.8. Матриця результатів алгоритму*

*авторська розробка

Наведені алгоритми визначення діабету є дуже точним і корисним для медичної практики. Проте, взаємодія з ними наразі може бути складною для медичних працівників, які не мають достатньої кваліфікації в галузі обробки даних та машинного навчання.

Таким чином сформуємо алгоритм, який проходить програма кожен раз на основі даних медичної установи.

В нас є доступним категоріальні дані та числові: Polyuria, Polydipsia, Gender, Sudden weight loss, Partial paresis, Polyphagia, Зважений коефіцієнт маси тіла (BMI), Гемоглобін A1C (HbA1c), Вік (age).

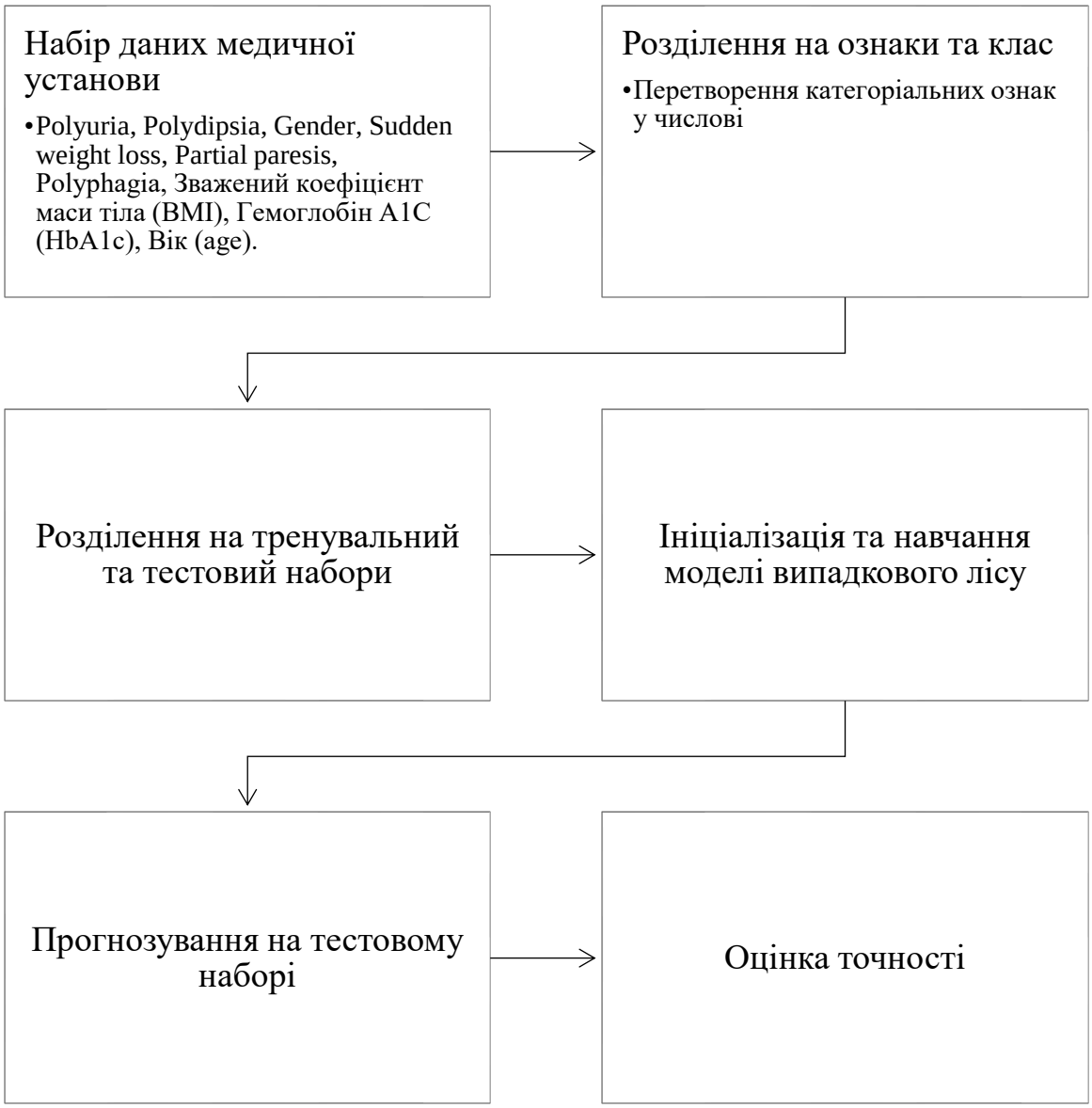

Рис.3.9 Алгоритм визначення діабету другого типу*

*авторська розробка

На першому етапі завантажуємо дані з медичної установи по встановленим ознакам. Ознаки (features) та цільова змінна (target) витягуються з датасету. Категоріальні ознаки перетворюються на бінарні. Далі дані розділяються на тренувальний та тестовий набори за допомогою train_test_split(). Модель випадкового лісу (Random Forest) ініціалізується, навчається на тренувальних даних і виконує прогнозування на тестових даних. Наостанку, оцінюється точність моделі з використанням метрики accuracy.

3.4. Висновки до третього розділу

У третьому розділі було зроблено значний крок у напрямку створення технології машинного навчання для визначення ДДТ. Була створена база знань, яка містить два набори даних з числовими та категоріальними значеннями і вихідну змінну клас. В результаті аналізу технічного стеку було вибрано мову програмування Python та відповідні бібліотеки машинного навчання.

Також було проведено тестування алгоритмів, які були проаналізовані у другому розділі, та виявлено, що метод випадкового лісу дає найкращі результати. Це дозволило проводити подальші дослідження з використанням цього методу.

Для покращення результатів було використано кореляційний аналіз для виокремлення найважливіших ознак у даних та метод головних компонент для категоріальних даних. Ці методи дозволили отримати більш точні результати визначення діабету на основі набору аналізів.

У підсумку, третій розділ дозволив сформувати технологію машинного навчання для визначення ДДТ, яка базується на точних методах та алгоритмах. Дана технологія може бути використана для подальшого дослідження та впровадження у медичній практиці.

РОЗДІЛ 4. ЗАСТОСУВАННЯ МОДЕЛІ ПРОГНОЗУВАННЯ ДДТ

4.1. Опис проекту застосування моделі прогнозування ДДТ

Маючи сформовану технологію визначення ДДТ методом випадкового лісу важливо зробити процес комунікації лікарів із нею зручним та зрозумілим.

Щоб зробити використання алгоритмів більш доступним і простим для лікарів, можна розробити застосунок з простим інтерфейсом та вбудованою аналітикою.

Пропонується інтеграція запропонованого алгоритму визначення ДДТ із Data Driven Ecosystem Microsoft. А саме використання зі сторони створення застосунку Power App, зі сторони збереження даних Azure Data Lake, зі сторони розгортання моделі машинного навчання Azure Machine Learning, зі сторони візуалізації даних Power BI.

Для створення проекту інтеграції запропонованого алгоритму визначення ДДТ з Data Driven Ecosystem Microsoft, необхідно виконати наступні кроки:

Створити Power App за допомогою сервісу Microsoft Power Apps. Power App повинен мати форму для введення даних пацієнта. Після введення даних, Power App повинен відправляти дані в Azure Data Lake для збереження.

Після збереження даних в Azure Data Lake, вони повинні бути доступні для обробки на Azure Machine Learning. Тут виконується завантаження даних, побудова моделі машинного навчання на підставі алгоритму визначення ДДТ і навчання цієї моделі.

Azure Synapse Analytics надає змогу робити обробку даних, чистити їх та підлаштовувати під модель.

Модель необхідно розгорнути на Azure Machine Learning. Для цього необхідно створити веб-сервіс, що надає доступ до моделі машинного навчання за допомогою API.

В Power App необхідно зробити запит до веб-сервісу на Azure Machine

Learning, щоб отримати результати прогнозування діабету для введених даних пацієнта.

Для візуалізації даних можна використовувати Power BI. Power BI дозволяє створювати різноманітні звіти і діаграми на основі даних, які зберігаються в Azure Data Lake.

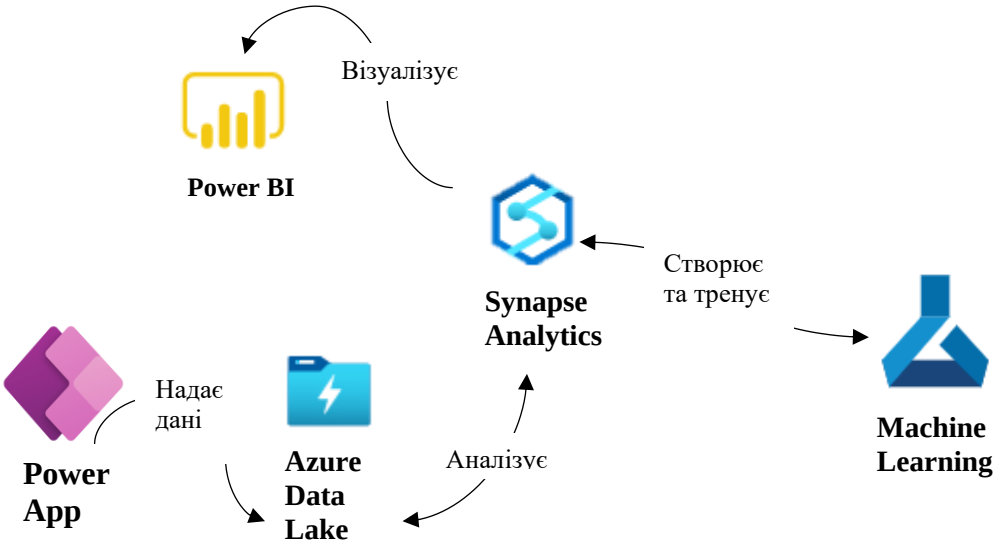

Рис.4.1. Архітектура рішення*

*авторська розробка

Інтеграція запропонованого алгоритму визначення ДДТ з Data Driven Ecosystem Microsoft має декілька переваг:

Ця система дозволяє зберігати дані пацієнтів в обліковому записі Azure Data Lake, що забезпечує безпеку та доступність даних для будь-якого пристрою, підключеного до Інтернету.

Використання Power App дозволяє лікарям та медичним працівникам зручно вносити дані пацієнтів та бачити результати прогнозування діабету з використанням зручного та простого інтерфейсу.

Розгортання моделі машинного навчання на Azure Machine Learning забезпечує високу точність та ефективність прогнозування діабету, що може значно поліпшити якість лікування та підвищити результативність медичної діяльності.

Використання Power BI для візуалізації даних дозволяє медичним працівникам легко переглядати та аналізувати дані про пацієнтів, що може сприяти прийняттю кращих та ефективніших рішень.

Рис. 4.2. SWOT аналіз проекту*

*авторська розробка

Після опрацювання сильних та слабких сторін проекту важливо позиціонувати можливості та загрози. Таблиця відображає дії та їх вірогідність

відбування з високою, середньою та слабкою імовірністю.

Таблиця 4.1

Матриця «вірогідність/вплив» для позиціювання можливостей зовнішнього середовища*

Дія	**Вірогідність**			
		Висока	**Середня**	**Слабка**
	Сильне	Державна підтримка проекту Поява високоточних моделей	Підвищення рівня освіти населення з питань діабету	Зацікавленість людей у покращення продукту і надання даних і зворотного зв'язку
	Помірне	Стратегія блакитного океану наддасть можливості вільно обирати інструменти т стратегії	Збільшення рівня доходів	Об'єднання з конкурентами у велику компанію
	Слабке	Відсутність достойних аналогії	Відсутність податків з боку держави, пільги	Підвищення рівня освіти людей з питань машинного навчання

*авторська розробка

Дія появи високоточних моделей може мати середню вірогідність підвищення рівня освіти населення з питань діабету, або слабку вірогідність підвищення рівня освіти людей з питань машинного навчання. За відсутності достойних аналогій дія може мати слабку вірогідність, а також підвищення рівня освіти людей з питань машинного навчання.

Наприклад, висока вірогідність державної підтримки проекту при сильній дії, що означає, що якщо буде зроблено зусилля, то цей проект може отримати значну підтримку від держави.

Таблиця 4.2

Матриця «вірогідність/вплив» для позиціювання погроз зовнішнього середовища*

	Вірогідність			
Дія		**Висока**	**Середня**	**Низька**
	Сильна	Зменшення рівня життя	Поява більш глобальних проблем, що відволікатимуть від проблем здоров'я	Лікарні абсолютно не зацікавлені у підвищенні своєї кваліфікації
	Помірна	Поява сильного конкурента, з високим рівнем довіри від споживачів	Можливі хакерські атаки	Помилки лікарів при введенні даних
	Слабка	Погіршення екології	Поява іноземних конкурентів	Зросте попит на альтернативну/народну медицину

*авторська розробка

Аналіз PEST — це інструмент вимірювання, який використовується для оцінки ринків для певного продукту чи бізнесу в певний період часу. PEST означає політичні, економічні, соціальні та технологічні фактори.

Основна мета PEST-аналізу — зрозуміти, які зовнішні сили можуть впливати на вашу організацію та як ці фактори можуть створювати можливості чи загрози вашому бізнесу. Аналіз PEST допомагає зрозуміти поточні зовнішні впливи на бізнес, аби працюватина фактах, а не на припущеннях.

Таким чином найбільш важливими факторами у реалізації продукту стали

соціальні та технологічні, адже першочергово люди є клієнтами та постачальниками у діяльності застосунку, а наразі рівень та спосіб життя, на жаль, лише провокує появу цукрового діабету.

Таблиця 4.3

PEST-аналіз галузі виявлення діабету методами машинного навчання*

Фактори	**Вплив**	**Оцінка впливу (1-5)** 5- найсильніше	**Ризик змін (0-100%)**
Політичні			
Бюрократія	Негативний	3	10
Закон про захист даних	Позитивний	1	50
Податкова політика (податкові ставки та пільги)	Негативний	3	10
Економічні			
Інфляція	Негативний	4	20
Процентні ставки	Негативний	3	20
Попит / Пропозиція	Позитивний	5	30
Соціальні			
Спосіб життя	Позитивний	5	40
Популяція	Позитивний	5	10
Свідомість щодо здоров’я	Позитивний	5	15
Технологічні			
Кібербезпека	Позитивний	5	60
Хмарні бази даних	Позитивний	5	60
Доступність інтернету (рівень покриття)	Негативний	5	40

*авторська розробка

Технологічні чинники, за умов їх покращення з часом створюють безпечну платформу для реалізації захищеного застосунку з доступом у кожному селі України. Маємо розуміти, що окрім основного роду діяльності – визначенні діабету за аналізами, діяльність має розвернутися й у освітній функції.

4.2. Розгортання проекту

Архітектурне рішення проекту включає в себе поєднання декількох сервісів та їх взаємодію. Для того щоб його розгорнути необхідно вибудувати команду та процеси.

В загальному проект матиме певні ітерації розробки.

Розробка Power App для внесення даних пацієнтів: потрібно розробити додаток, який дозволить лікарям ввести дані пацієнтів та отримати звітність в першу чергу. Пізніше можна дороблювати інші модулі. Введені дані можна зберегти у форматі JSON, щоб їх можна було легко відправити з Azure Data Lake Storage.

Інтеграція з Azure Data Lake Storage: Після збереження даних потрібно створити контейнер в Azure Data Lake Storage та налаштувати з'єднання з додатком. Для з'єднання з Azure Data Lake Storage можна використовувати Python SDK.

Модель машинного навчання в Azure Machine Learning: треба розгорнути запропоновану модель як веб сервіс.

Для того, щоб отримати прогнози, потрібно створити API, який буде взаємодіяти з моделлю машинного навчання. Це можна зробити за допомогою Azure Machine Learning SDK.

Далі треба розмістити API на Azure Container Instances або Azure Kubernetes Service. Це дозволить запускати API в хмарі і надавати доступ до нього з будь-якого місця.

Після того, як дані були внесені, додаток повинен звертатися до API, щоб отримати прогнози, і повернути їх до лікаря. Та зберегти їх у базі даних.

Power BI підключається до бази даних та використовує їх. Потів звіт публікується у хмару та вбудовується у додаток.

Для якісної імплементації даного проекту необхідна команда з певною організацією. Наразі більш сучасними та ефективними є гнучкі (органічні)

організаційні структури.

Органічні структури відомі своїм широким діапазоном контролю, децентралізацією, низькою спеціалізацією. Ця модель є набагато менш формальною, ніж механістична, забезпечує гнучкість, що може бути надзвичайно корисно для бізнесу, який орієнтується в швидкій галузі або просто намагається стабілізувати себе після важкого кварталу. Це також дає співробітникам можливість пробувати нове та розвиватися як професіонали, що робить робочу силу організації більш потужною в довгостроковій перспективі. Стартапи часто ідеально підходять для органічної структури, оскільки вони просто намагаються здобути пізнаваність бренду.

Розгортання проекту з використанням Data Driven Ecosystem Microsoft потребує залучення різних спеціалістів. Нижче наведено основний склад команди, яка може реалізувати такий проект:

Менеджер проекту - відповідає за керування всіма етапами проекту, встановлює зв'язки між різними членами команди, забезпечує зв'язок з клієнтом, встановлює терміни і забезпечує виконання проекту відповідно до умов договору.

Дослідник даних - відповідає за збір, очищення і підготовку даних для відображення на звітах.

Спеціаліст з машинного навчання - відповідає за розробку і вдосконалення алгоритмів машинного навчання для визначення діабету, побудови і навчання моделі машинного навчання, а також за її тестування і оптимізацію.

Архітектор даних - відповідає за проектування бази даних для зберігання та обробки даних в режимі реального часу.

Розробник Power App - відповідає за розробку функціоналу Power App для внесення даних користувачами, взаємодії з базою даних та відображення результатів.

Дизайнер – відповідає за створення зручного та гарного інтерфейсу застосунку та дизайнів звітів.

Системний адміністратор - відповідає за налагодження та підтримку

інфраструктури Azure, забезпечення безпеки даних, налагодження взаємодії між різними елементами системи.

Тестувальник – відповідає за безперебійну роботу застосунку.

Таблиця 4.4

Потенційний склад команди*

Посада	**Опис**	**Заробітна плата**
Менеджер проекту	відповідає за керування всіма етапами проекту, встановлює зв'язки між різними членами команди, забезпечує зв'язок з клієнтом, встановлює терміни і забезпечує виконання проекту відповідно до умов договору.	$3500
Дослідник даних	відповідає за збір, очищення і підготовку даних для відображення на звітах.	$3000
Дизайнер	відповідає за створення зручного та гарного інтерфейсу застосунку та дизайнів звітів.	$2500
Розробник Power App	відповідає за розробку функціоналу Power App для внесення даних користувачами, взаємодії з базою даних та відображення результатів.	$3000
Системний адміністратор	відповідає за налагодження та підтримку інфраструктури Azure, забезпечення безпеки даних, налагодження взаємодії між різними елементами системи.	$4000
Архітектор даних	відповідає за проектування бази даних для зберігання та обробки даних в режимі реального часу.	$3500
Спеціаліст з машинного навчання	створює, підтримує та оптимізує модель	$3500
Power BI розробник	створює звіти	$2500
Тестувальник	тестує проект	$1500

*авторська розробка

Методологія Scrum є досить популярною в галузі розробки програмного забезпечення та дозволяє ефективно керувати проектами з багатьма змінними.

Для вашого проекту, що включає розробку програмного забезпечення, рекомендується використовувати методологію Scrum з наступних причин:

Ітераційний підхід: Scrum заснований на ітераційному підході, який дозволяє зосередитися на здійсненні більш коротких ітерацій, що сприяє швидкому виявленню та виправленню помилок.

Прозорість: Scrum забезпечує прозорість у проекті, яка дозволяє всім членам команди бути в курсі всіх дій та прийняти своєчасні рішення.

Адаптивність: Scrum дозволяє швидко реагувати на зміни та адаптуватися до нових вимог клієнта та ринку.

Керування ризиками: Scrum дозволяє виявляти та керувати ризиками на ранніх етапах проекту, що зменшує ризик невдачі та збільшує ймовірність успіху проекту.

Колективна власність: Scrum дозволяє команді приймати колективні рішення та бути власником проекту, що стимулює більш активну участь та залученість кожного члена команди.

Прикладом історії користувача для даного проекту може бути:

Історія користувача: як користувач, я хочу отримувати персоналізовані рекомендації щодо лікування діабету.
Опис: як користувач застосунку я хочу отримувати персоналізовані рекомендації щодо лікування діабету на основі моїх індивідуальних факторів ризику та історії хвороби
Критерії приймання: - Розробіть модель машинного навчання, яка може надавати персоналізовані рекомендації щодо лікування діабету. - Навчіть модель, використовуючи набір даних, який містить відповідні дані пацієнтів та інформацію про лікування діабету. - Переконайтеся, що модель легко інтегрується в застосунок і доступна для користувачів. - Переконайтеся, що модель є масштабованою та може обробляти великі обсяги даних. - Надайте користувачам чіткі та дієві рекомендації щодо лікування діабету на основі їхніх індивідуальних факторів ризику та історії здоров’я.

Рис.4.3. Приклад користувацької історії*

*авторська розробка

Для даного проекту зацікавленими сторонами є: користувачі, держава, інвестори, замовники, команда проекту, медичні заклади, ЗМІ.

Проведемо аналіз стейкхолдерів за методом Power/Interest Grid.

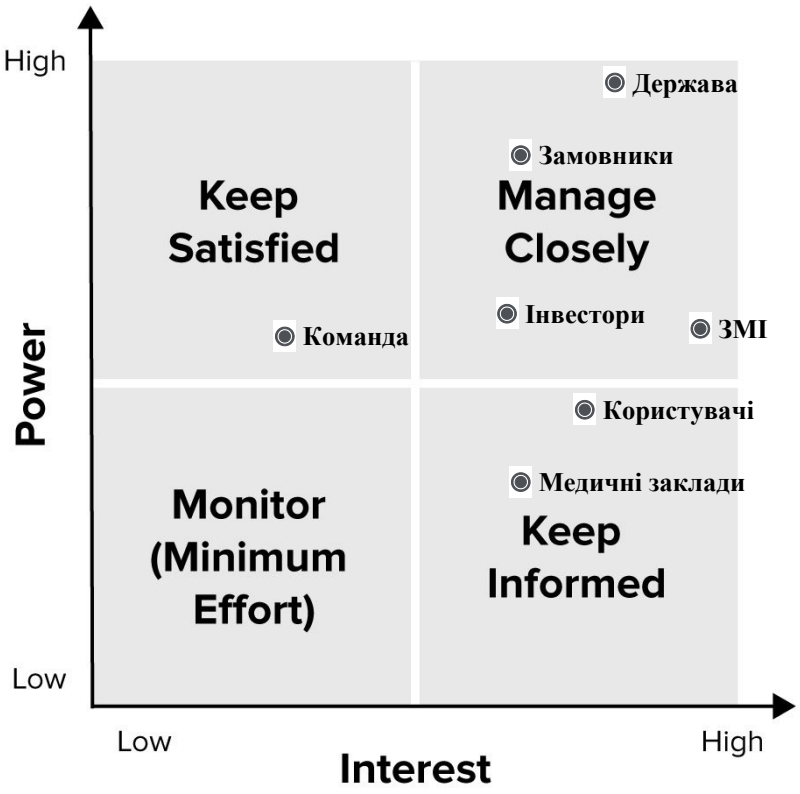

Рис.4.4. Power/Interest Grid для зацікавлених сторін проекту*

*авторська розробка

Таким чином до квадрату із сильним впливом на проект та сильною зацікавленістю у проекті потрапили держава, замовники, інвестори та засоби масової інформації. Держава має найбільший вплив, адже регулює медичну сферу та законодавство про медичні дані. В свою чергу замовники та інвестори можуть впливати на суть використання застосунку та його розвиток, а також зацікавлені в отриманні прибутку від роботи застосунку.

Квадрат низької зацікавленості та високого рівня впливу включає команду проекту, адже фактично команда зацікавлена у розвитку проекту на певному рівні, який визначатиме їх компенсацію та умови праці. Вони безпосередньо впливають на якість застосунку, а отже необхідно тримати команду задоволеною своїм робочим місцем.

До квадрату низької сили але великої зацікавленості відносимо користувачів та медичні заклади, як цільових клієнтів застосунку. Хочу вони найбільше користуються послугами застосунка, впливати на нього вони можуть через попит або мінорні зауваження.

Таблиця 4.5

Аналіз зацікавлених сторін*

№	Зацікавлені сторони	Вплив ЗС на проект	Вплив результатів проекту на ЗС
1	Держава	Нормативні акти, податкові знижки	Підвищення рівня освіти населення, визначення діабету на ранній стадії у населення, здоровіша нація
2	Замовники	Визначення обсягу та варіативності визначення хвороб	Підвищення конкурентоспроможності, підвищення доходів
3	Інвестори	Покращення фінансового стану, підвищення швидкості запуску проекту	Отримання частини доходів проекту, участь у соціально значущій роботі
4	ЗМІ	Представлення продукту проекту у населення	Поява нової ніші для відслідковування, поява нових партнерів
5	Команда	Якість отриманого продукту	Отримання компенсації, іміджу
6	Користувачі	Визначення найбільш доступних аналізів, зміни продукту на користь попиту	Покращення рівня освіти та життя
7	Медичні заклади	Визначають попит на продукт та його можливості	Підвищення загального рівня надання медичних послуг

*авторська розробка

4.3. Опис структурних елементів проекту

Серед найважливіших елементів проекту є Power App застосунок, модель машинного навчання, база даних, звітність у Power BI.

База даних проекту має бути побудована за принципами правильної побудови та підлаштована під використання у системи надання звітності.

Для цього мають бути присутні таблиці фактів та таблиці описів.

За допомогою сервісу Mockaroo було змодельовані наповнену базу даних проекту (орієнтовну).

Пропонується створення однієї факт таблиці та 4 описових таблиць.

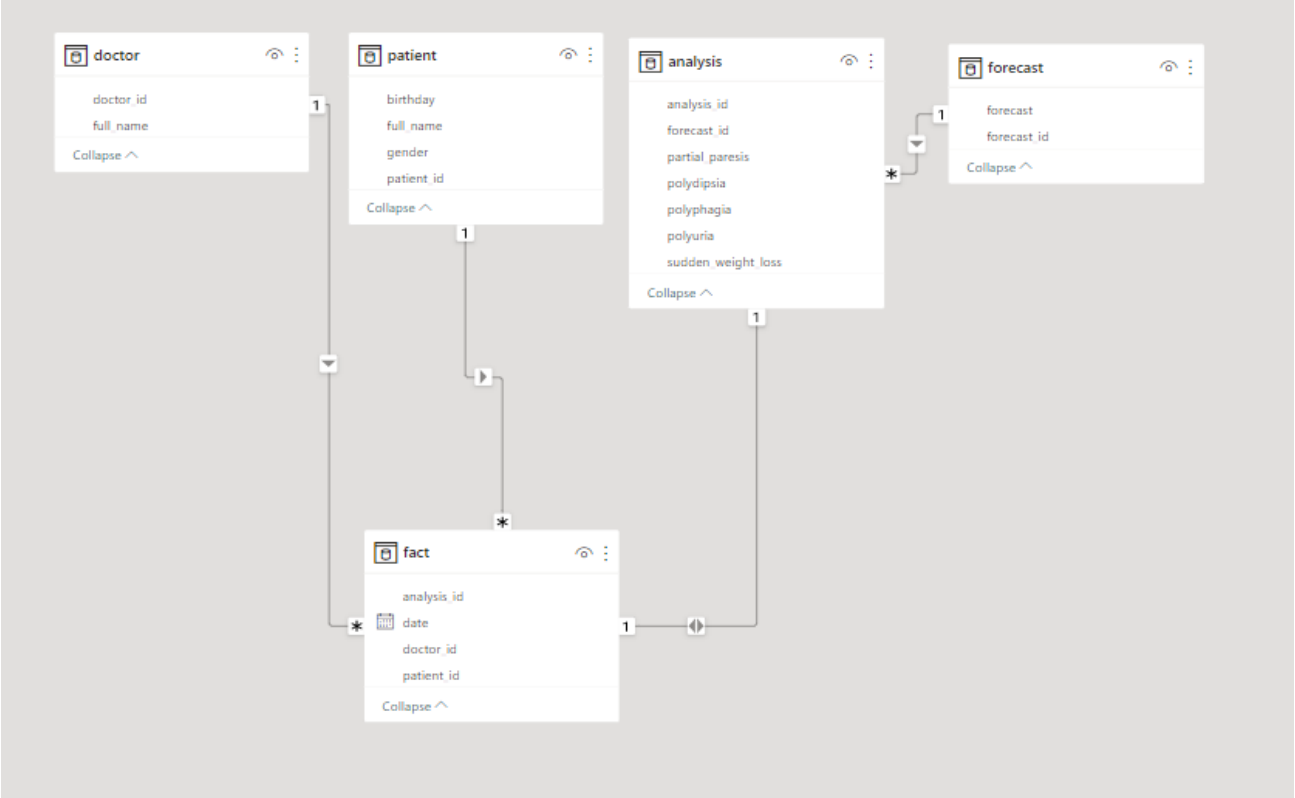

Рис. 4.5. Схема зірка для бази даних*

*авторська розробка

У факт таблицю має збиратися інформація за конкретним пацієнтом, його набором аналізів в певний проміжок часу та його прогноз. Описуючи таблиці містять інформацію про імена лікарів, пацієнтів, стать та вік, також аналізи та тип пронозу.

В загальному схема використовує принципи зірки, хоча і має одне відгалуження в прогнозі, але це є оптимізаційне рішення, адже туди має записуватися результат відпрацювання моделі машинного навчання. Детальніше можна ознайомитися у додатку В.

Зіркова схема в DWH дозволяє легко моделювати аналітичні структури даних, такі як фактичні таблиці та вимірювані таблиці. Це забезпечує швидкий доступ до даних та ефективні операції агрегації та фільтрації. Зіркова схема

дозволяє легко додавати нові факти та вимірювання до моделі даних. Це дозволяє розширювати функціональність аналітичної системи та пристосовувати її до змінних потреб бізнесу.

На тестових даних пропонується зразок звіту, який може бути інтегрований у застосунок.

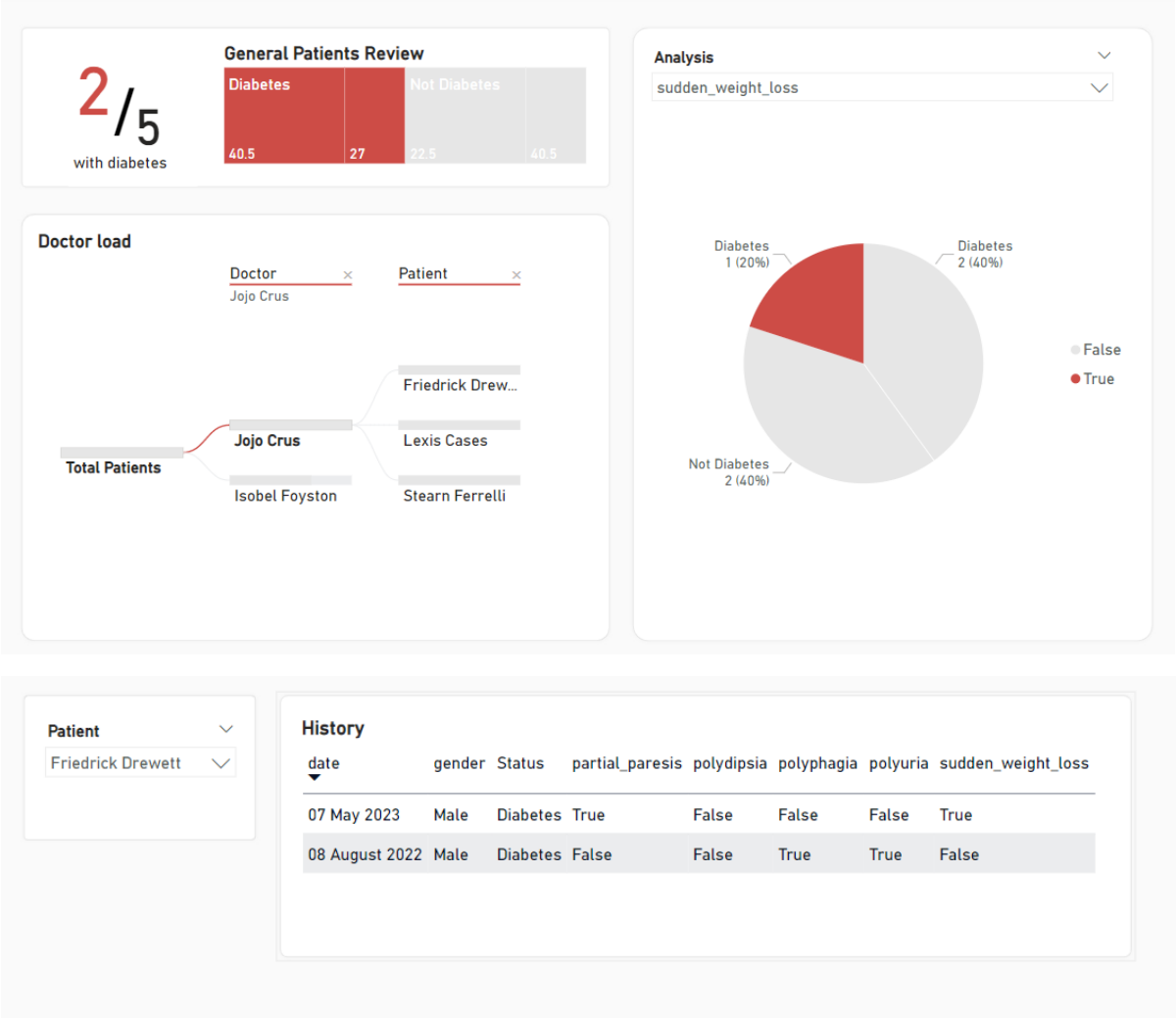

Рис.4.6. Зразок звітності для застосунку*

*авторська розробка

Пропонується надавати лікарям інформацію в загальному по пацієнтам, під їх підпорядкуванням, а також виокремлено по аналізам виключного пацієнта.

Водночас це лише зразок звіту, на практичній реалізації пропонується створити декілька звітів та розширити зразок.

- Графіки та діаграми для візуалізації даних про пацієнтів, таких як стать, вік та рівень глюкози в крові.

- Загальна статистика щодо кількості пацієнтів, що мають ДДТ, та їх відсоткового співвідношення до загальної кількості пацієнтів.
- Графіки та діаграми, що відображають залежність рівня глюкози в крові від інших факторів, таких як вік, стать, артеріальний тиск, індекс маси тіла.
- Аналіз того, які фактори є найбільш важливими для визначення наявності ДДТ.
- Звіт про ефективність моделі машинного навчання, що визначає ДДТ, включаючи точність та показники confusion matrix.
- Звіт про результати тестування моделі на нових даних та порівняння її ефективності з іншими моделями.
- Аналіз того, як змінні впливають на результати моделі та можливі шляхи їх оптимізації.

Наступним не менш важливим елементом є створення Power App застосунку.

Power App застосунок для проекту з визначення діабету повинен містити наступні елементи:

- Форма введення даних пацієнта, включаючи ім'я, стать, дату народження та інші характеристики, необхідні для діагностики діабету.
- Можливість завантаження результатів аналізів та лікарських призначень пацієнта.
- Інтерфейс для перегляду результатів діагностики та рекомендацій щодо лікування.
- Функція зв'язку з лікарем або медичним закладом для консультації та отримання додаткової інформації.
- Опція збереження даних пацієнта та їх зручного перегляду.
- Функціонал для надання користувачеві порад з питань життєвого стилю та харчування для підтримки здорового способу життя.
- Можливість підключення до сервісів Azure для зберігання та обробки даних.

- Функціонал для аналізу та візуалізації даних з метою виявлення тенденцій та покращення процесу діагностики та лікування.
- Захист даних та можливість керування доступом до них.
- Функціонал для забезпечення розширення функціональності застосунку в майбутньому з використанням зовнішніх компонентів та API.

Додаток торкається чутливих медичних даних та має забезпечити високий рівень безпеки даних.

Для забезпечення безпеки застосунку можна вжити декілька заходів та кращих практик:

- Аутентифікація та авторизація: механізми аутентифікації, щоб перевіряти ідентичність користувачів, які мають доступ до застосунку. Встановлення рівнів доступу та прав користувачів, щоб обмежити доступ до конфіденційної інформації.
- Захист даних: шифрування для захисту конфіденційної інформації, яка передається між застосунком та сервером.
- Запобігання атакам: заходи для запобігання атакам, таким як вразливості перехоплення, зловмисного коду та SQL-ін'єкції. Використання фільтрації введення та валідацію даних, а також захист від перебору паролів.
- Моніторинг: активності користувачів, події та помилки для подальшого аналізу та виявлення потенційних проблем.

Потенційними сценаріям застосування застосунку в реальному житті є:

- Мобільний додаток для самомоніторингу: додаток може бути використаний як той, що надає результат на основі даних з роботи для самоідентифікації, де доступ до лікарів є обмеженим, який дозволяє користувачам вводити свої показники, такі як рівень глюкози в крові, коефіцієнт маси тіла і інші симптоми. Додаток може використовувати модель для надання користувачам ризикових оцінок та порад щодо контролю діабету.
- Планування лікування та догляду в медичній установі: Використання моделі для розробки системи планування лікування та догляду для пацієнтів з

діабетом другого типу в межах певної установи. Система може аналізувати історію пацієнта та прогнозувати оптимальний план лікування, включаючи призначення ліків, дієтичні рекомендації та фізичну активність. Застосунок може бути використаний для розробки системи попередження ускладнень діабету другого типу. Вона може моніторити показники пацієнтів та виявляти сигнали ризику ускладнень, таких як гіперглікемія або гіпоглікемія. Пацієнти та медичний персонал можуть отримувати сповіщення та поради щодо подальших кроків.

В запропонованому додатку є обмеження:

- Залежність від введених даних: Для точності прогнозування додаток потребує правильного та достовірного введення показників та симптомів. Некоректне введення або неправильні дані можуть призвести до неточних результатів.
- Відсутність медичної консультації: Додаток може надавати інформацію та рекомендації, але не замінює професійну медичну консультацію. Важливо пам'ятати, що діагноз та лікування діабету повинні здійснюватися під наглядом кваліфікованого медичного фахівця.

Можливості:

- Самомоніторинг та управління: Додаток може допомогти користувачам вести самомоніторинг своїх показників та симптомів, таких як рівень глюкози в крові, вага та фізична активність. Він може надавати сповіщення та поради щодо контролю діабету, сприяючи здоровому способу життя.
- Зручний доступ до інформації: Додаток може надати користувачам зручний та швидкий доступ до інформації про діабет другого типу, його симптоми, лікування та попередження ускладнень. Це дозволяє пацієнтам бути освіченими та керувати своїм станом здоров'я.
- Моніторинг трендів та статистики: Додаток може відстежувати та аналізувати дані користувачів для виявлення трендів та статистики щодо контролю діабету.

4.4. Висновки до четвертого розділу

У четвертому розділі було запропоновано приклад реалізації практичного застосування запропонованої технології визначення ДДТ методом машинного навчання випадкового лісу на числових та категоріальних даних.

Було запропонована підхід Data Driven Ecosystem Microsoft з використанням його елементів, таких як Data Lake Storage, Power BI, Power App, Machine Learning.

Усі вони в поєднанні надають можливість розгорнути запропоновану технологію визначення ДДТ у хмарі та використовувати її лікарями у зручному форматі. Також запропонована архітектура дає можливість отримувати аналітичні продукти.

Було також запропонована структура проекту, методологія та наведено приклад користувацької історії.

Було побудовано базу даних у вигляді зірки з тестовими даними, а також побудовано звіт на основі структури даних, який в подальшому можна застосовувати на реальних даних для отримання загальної статистики та окремо по пацієнтах певного лікаря.

ВИСНОВКИ

Таким чином було досягнуто таких результатів.

Було детально розглянуто предметну область захворювання ДДТ та його визначення. З метою збільшення розуміння та глибшого аналізу проблеми, була надана детальна характеристика захворювання, що охоплювала як його причини, так і можливості виявлення. Важливою складовою цього дослідження було визначення трендів розвитку захворювання та попиту у населення стосовно цього питання. Було з’ясовано, що ДДТ стає все більш поширеним у світі та є одним з найбільш впливових захворювань на здоров’я населення. Далі, була проведена аналітична робота зі світової та української практики визначення ДДТ, під час якої було розглянуто декілька рішень та знайдено їх переваги та недоліки. Було виявлено, що на сьогоднішній день в Україні немає конкурентоспроможного рішення для визначення ДДТ, тому потреба в такому рішенні є нагальною.

Після цього, було з’ясовано, що одного лікарського висновку не є достатньо для точного та надійного визначення ДДТ. Тому, для підвищення ефективності визначення цього захворювання, було запропоновано використати ММН. Ці методи дозволяють зібрати та обробити великі обсяги даних, що дозволяє забезпечити точність та надійність діагностики.

Отже, була поставлена задача для вирішення проблеми низького та неточного визначення ДДТ у створенні технології визначення за допомогою машинного навчання.

Було проаналізовано різні алгоритми машинного навчання, їх класифікацію та особливості застосування. Було надано математичний опис методів машинного навчання, а саме логістична регресія, SVM та KNN методи, та метод випадково лісу. Проведено комплексний аналіз переваг та недоліків кожного з методів. Опрацьовано актуальний стан методів імплементації методів.

Також було надано формули та приклади використання кожного з цих методів. Наприклад, була надана формула для логістичної регресії та описано її

застосування для бінарної класифікації.

Після того, як було проаналізовано різні ММН, було проведено дослідження статей, в яких використовувалися ці методи. Зокрема, було проаналізовано статті, що стосуються визначення ДДТ.

В результаті аналізу було обрано метод випадкового лісу (Random Forest) як найбільш оптимальний метод для вирішення задачі визначення ДДТ. Це пов'язано з тим, що метод Random Forest відносно простий у реалізації, дозволяє враховувати велику кількість ознак та має досить високу точність прогнозування.

Зроблено значний крок у напрямку створення технології машинного навчання для визначення ДДТ. Була створена база знань, яка містить два набори даних з числовими та категоріальними значеннями і вихідну змінну клас. В результаті аналізу технічного стеку було вибрано мову програмування Python та відповідні бібліотеки машинного навчання.

Також було проведено тестування алгоритмів, які були проаналізовані у другому розділі, та виявлено, що метод випадкового лісу дає найкращі результати. Це дозволило проводити подальші дослідження з використанням цього методу.

Для покращення результатів було використано кореляційний аналіз для виокремлення найважливіших ознак у даних та метод головних компонент для категоріальних даних. Ці методи дозволили отримати більш точні результати визначення діабету на основі набору аналізів.

У підсумку, сформовано технологію машинного навчання для визначення ДДТ, яка базується на точних методах та алгоритмах. Дана технологія може бути використана для подальшого дослідження та впровадження у медичній практиці. У четвертому розділі було запропоновано приклад реалізації практичного застосування запропонованої технології визначення ДДТ методом машинного навчання випадкового лісу на числових та категоріальних даних.

Було запропонована підхід Data Driven Ecosystem Microsoft з використанням його елементів, таких як Data Lake Storage, Power BI, Power App,

Machine Learning.

Усі вони в поєднанні надають можливість розгорнути запропоновану технологію визначення ДДТ у хмарі та використовувати її лікарями у зручному форматі. Також запропонована архітектура дає можливість отримувати аналітичні продукти.

Було також запропонована структура проекту, методологія та наведено приклад користувацької історії.

Було побудовано базу даних у вигляді зірки з тестовими даними, а також побудовано звіт на основі структури даних, який в подальшому можна застосовувати на реальних даних для отримання загальної статистики та окремо по пацієнтах певного лікаря.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1 AACE/ACE Guidelines for the Management of Dyslipidemia and Prevention of Cardiovascular Disease Writing Committee, Endocr Pract. 2017;23(Suppl 2).

2 Aditya Sharma. Principal Component Analysis (PCA) in Python .. URL.: https://www.datacamp.com/tutorial/principal-component-analysis-in-python

3 American Diabetes Association. Economic costs of diabetes in the US in 2017. Diabetes Care. 2018 May;41(5):917–928.

4 American Diabetes Association. Standards of Medical Care in Diabetes—2021. Diabetes Care. 2021 Jan 1; 44 (Supplement 1).

5 Analysis Of Decision Tree For Diabetes Prediction. URL.: https://www.researchgate.net/publication/334969613_Analysis_Of_Decision_Tree_For_Diabetes_Prediction.

6 Artificial Neural Networks Model for Predicting Type 2 Diabetes Mellitus Based on VDR Gene FokI Polymorphism, Lipid Profile and Demographic Data. URL.: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7465516/.

7 Ayush Pant. Introduction to Logistic Regression . / Ayush Pant. 2019. URL: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148.

8 Cardiogram: Early Detection and Prevention. URL.: https://ilumivu.com/products/cardiogram/.

9 Centers for Disease Control and Prevention. National Center for Health Statistics. About Underlying Cause of Death 1999–2019; CDC WONDER Online Database. Accessed at http://wonder.cdc.gov/ucd-icd10.html on Sept 17, 2021.

10 Comparing Different Programming Languages For Machine Learning. 2020. URL: https://analyticsindiamag.com/comparing-different-programming-languages-for-machine-lea

11 Top 9 Python Libraries for Machine Learning in 2021. URL: https://www.upgrad.com/blog/top-python-libraries-for-machine-learning/.rning/.

12 Convolutional Neural Network for Manufacturing Platform. URL.: https://www.infopulse.com/case-studies/implementation-of-convolutional-neural-network-for-manufacturing-needs
13 COVID-19 Access Control: an eID access control solution (MVP). URL.: https://www.infopulse.com/case-studies/covid-19-access-control
14 Deberneh HM, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int J Environ Res Public Health. 2021 Mar 23;18(6):3317. doi: 10.3390/ijerph18063317. PMID: 33806973; PMCID: PMC8004981.
15 Decision tree with the IRIS dataset in Python. URL.: https://www.educative.io/edpresso/how-to-build-a-decision-tree-with-the-iris-dataset-in-python
16 Diabetes Diagnostic Prediction Using Vector Support Machines. URL.: https://www.sciencedirect.com/science/article/pii/S1877050920305020.
17 Diabetes Prediction using Machine Learning Algorithms. URL.: https://www.researchgate.net/publication/339543101_Diabetes_Prediction_using_Machine_Learning_Algorithms.
18 Diabetes statistics 2021 By SingleCare Team. URL.: https://www.singlecare.com/blog/news/diabetes-statistics/#:~:text=In%201980,%20108%20million%20people,will%20have%20diabetes%20by%202045.
19 Digital Platform for Temporary Staff Placing & Scheduling. URL.: https://www.infopulse.com/case-studies/healthcare-staff-marketplace
20 Divers J, Mayer-Davis EJ, Lawrence JM, et al. Trends in Incidence of Type 1 and Type 2 Diabetes Among Youths— Selected Counties and Indian Reservations, United States, 2002–2015. MMWR Morb Mortal Wkly Rep. 2020 Feb 14;69(6):161–165.
21 DreaMed Diabetes: Diabetes AI solutions. URL.: https://www.google.com/search?q=DreaMed+Diabetes.
22 Dutta, A.; Hasan, M.K.; Ahmad, M.; Awal, M.A.; Islam, M.A.; Masud, M.; Meshref, H. Early Prediction of Diabetes Using an Ensemble of Machine Learning

Models. Int. J. Environ. Res. Public Health 2022, 19, 12378. https://doi.org/10.3390/ijerph191912378

23 Dwivedi, Karnika & Sharan, Hari & Vishwakarma, Vinod. (2019). Analysis Of Decision Tree For Diabetes Prediction. International Journal of Engineering and Technical Research (IJETR). 9. 10.31873/IJETR.9.6.2019.64.

24 Fedirko Yu., Yehorchenkov O. Deep Learning Model in Predicting Diabetes (Type 2) Information Technology and Implementation (Satellite): Conference Proceedings, December 01, 2022, Kyiv, Ukraine / Taras Shevchenko National University of Kyiv and [etc]; Vitaliy Snytyuk (Editor). – Kyiv: Publisher Individual entrepreneur Picha Y.V., 2022. 186 p.

25 Fedirko Yu., Yehorchenkov O. Random Forest Algorithm in Predicting Diabetes (Type 2) Information Technology and Implementation (Satellite): Conference Proceedings, December 02, 2021, Kyiv, Ukraine / Taras Shevchenko National University of Kyiv and [etc]; Vitaliy Snytyuk (Editor). – Kyiv: Publisher Individual entrepreneur Picha Y.V., 2021. 185 p.

26 G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 1009-1014, doi: 10.1109/ICRITO48877.2020.9197832.

27 Gas Turbine Monitoring using a Digital Twin Data Platform. URL.: https://www.infopulse.com/case-studies/gas-turbine-monitoring-anomaly-detection

28 Hunter Heidenreich. What are the types of machine learning? . / Hunter Heidenreich. 2018. URL: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f.

29 IDF Diabetes Atlas | Tenth Edition. URL.: https://diabetesatlas.org

30 Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74, 2204–2214. doi: 10.2307/1939574

31 Jobeda Jamal Khanam, Simon Y. Foo, A comparison of machine learning algorithms for diabetes prediction, ICT Express, 2021, ISSN 2405-9595, https://doi.org/10.1016/j.icte.2021.02.004.

32 Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. R. News 2, 18–22.

33 M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic", Journal of Intelligent Learning Systems and Applications, 2017.

34 One Drop: Best Diabetes Management System. URL.: https://onedrop.today.

35 Onesmus Mbaabu. Introduction to Random Forest in Machine. 2020. URL.: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

36 Poly TN, Islam MM, Li YJ. Early Diabetes Prediction: A Comparative Study Using Machine Learning Techniques. Stud Health Technol Inform. 2022 Jun 29;295:409-413. doi: 10.3233/SHTI220752. PMID: 35773898.

37 Pranto, Md Badiuzzaman & Mehnaz, Sk & Mahid, Esha Bintee & Sadman, Imran & Rahman, Ahsanur & Momen, Sifat. (2020). Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh. Information. 11. 374. 10.3390/info11080374.

38 Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. URL.: https://pubmed.ncbi.nlm.nih.gov/31826018/.

39 Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. URL.: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8306487/.

40 Prediction Model of Type 2 Diabetes Mellitus for Oman Prediabetes Patients Using Artificial Neural Network and Six Machine Learning Classifiers. URL.: https://www.mdpi.com/2076-3417/13/4/2344.

41 Prediction of diabetes using logistic regression and ensemble techniques. URL.: https://www.sciencedirect.com/science/article/pii/S2666990021000318.

42 Prediction of Type 2 Diabetes using Machine Learning Classification Methods. URL.:

https://www.researchgate.net/publication/340700070_Prediction_of_Type_2_Diabetes_using_Machine_Learning_Classification_Methods.

43 Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques. URL.: https://www.frontiersin.org/articles/10.3389/fcomp.2022.835242/full.

44 Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. URL.: https://nutrition.bmj.com/content/early/2021/03/09/bmjnph-2020-000200.

45 Rashid, Ahlam (2020), “Diabetes Dataset”, Mendeley Data, V1, doi: 10.17632/wj9rwkp9c2.1

46 Reducing Production Defects with Predictive Quality Models. URL.: https://www.infopulse.com/case-studies/predictive-quality-models-and-advanced-analytics-reduce-production-defects-and-expenses

47 Risk prediction of type II diabetes based on random forest model. URL.: https://ieeexplore.ieee.org/document/7972337.

48 Sajida Perveena, Muhammad Shahbaza, Aziz Guergachib and Karim Keshavjeec, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Procedia Computer Science, vol. 82, pp. 115-121, 2106.

49 Saniya Parveez, Roberto Iriondo. Principal Component Analysis (PCA) with Python .. URL.: https://pub.towardsai.net/principal-component-analysis-pca-with-python-examples-tutorial-67a917bae9aa

50 Scalable RPA Solution for the Healthcare Institution of North Iceland . URL.: https://www.infopulse.com/case-studies/scalable-rpa-solution-hsn-power-automate

51 Significance of machine learning in healthcare: Features, pillars and applications. URL.: https://www.sciencedirect.com/science/article/pii/S2666603022000069.

52 Su X, Kong Y, Peng D. Evidence for changing lipid management strategy to focus on non-high density lipoprotein cholesterol. Lipids Health Dis. 2019 Jun 7;18(1):134.

53 Sunil Gupta, Kamal Saluja, Ankur Goyal, Amit Vajpayee, Vipin Tiwari, Comparing the performance of machine learning algorithms using estimated accuracy, Measurement: Sensors, Volume 24, 2022, 100432, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2022.100432. (https://www.sciencedirect.com/science/article/pii/S2665917422000666)

54 Sweetch-Health. URL.: https://www.sweetch.com.

55 Symptoms & Causes of Diabetes by U.S. Department of Health and Human Services. URL.: https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes

56 Taber, J. M., Leyva, B., & Persoskie, A. (2015). Why do people avoid medical care? A qualitative study using national data. Journal of general internal medicine, 30(3), 290–297. https://doi.org/10.1007/s11606-014-3089-1

57 Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain, Muhammad Awais Malik, Muhammad Mehdi Raza, Salman Ibrar, Zunish Abbas,A model for early prediction of diabetes,Informatics in Medicine Unlocked,Volume 16,2019,100204,ISSN 2352-9148,https://doi.org/10.1016/j.imu.2019.100204.

58 The Math Behind KNN .. URL: https://ai.plainenglish.io/the-math-behind-knn-7883aa8e314c.

59 Type 2 Diabetes Prediction Using Machine Learning Algorithms. URL.: https://jorjanijournal.goums.ac.ir/article-1-738-en.pdf.

60 Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier. URL.: https://www.sciencedirect.com/science/article/pii/S2666307420300073.

61 Type-II Diabetes Prediction Using Machine Learning Algorithms. URL.: https://ieeexplore.ieee.org/document/9740844.

62 Understanding the mathematics behind Support Vector Machines .. URL: https://shuzhanfan.github.io/2018/05/understanding-mathematics-behind-support-vector-machines/.

63 Usman Malik. Principal Component Analysis (PCA) in Python with Scikit-Learn .. URL.: https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/

64 What is the best programming language for Machine Learning .. 2017. URL: https://towardsdatascience.com/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7.

65 Yildirim, D., Küçüktopcu, E., Cemek, B. et al. Comparison of machine learning techniques and spatial distribution of daily reference evapotranspiration in Türkiye. Appl Water Sci 13, 107 (2023). https://doi.org/10.1007/s13201-023-01912-7

66 Zhovtukhin D., Yehorchenkov O. Classification of Bottles Images Using Convolutional Neural Networks .. URL: http://iti.fit.univ.kiev.ua/wp-content/uploads/ITI_Satellite_2020_Conference-Proceedings.pdf.

67 Басараб, М. Р., Іванько, Е. О., & Кулкарні, В. (2021). Прогнозування розвитку гестаційного цукрового діабету у вагітних із використанням методів машинного навчання. Мікросистеми, Електроніка та Акустика, 26(2), 228845–1 . https://doi.org/10.20535/2523-4455.mea.228845

68 Гугл тренди щодо пошуків причин цукрового діабету .. 2020. URL.: https://trends.google.ru/trends/explore?date=all&q=diabetes%20symptoms

69 Документація бібліотеки. URL.: https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart

70 Моделі подання знань. Мережеві моделі: фрейми, семантичні мережі .. URL.: http://baklaniv.at.ua/PSAI/lekcija_9-10_2016.2.pdf

71 Набір даних. URL.: https://archive.ics.uci.edu/ml/machine-learning-databases/00529/diabetes_data_upload.csv

72 Офіційний сайт компанії ТОВ Інфопульс. URL.: https://www.infopulse.com.

73 Статистика щодо цукрового діабету: чому Україні наразі необхідні об’єктивні дані? .. 2020. URL: https://www.apteka.ua/article/572594.

ДОДАТКИ

Додаток А

Назва датасету: Diabetes Dataset

Автор: Rashid, Ahlam

Рік публікації: 2020

Опис змінних:

- Age: Вік пацієнта.
- Gender: Стать пацієнта (чоловік або жінка).
- Polyuria: Наявність поліурії (збільшення кількості сечі).
- Polydipsia: Наявність полідипсії (надмірна спрага).
- Sudden weight loss: Наглий втрату ваги.
- Weakness: Слабкість.
- Polyphagia: Наявність поліфагії (надмірний апетит).
- Genital thrush: Наявність генітального трішини (грибкова інфекція).
- Visual blurring: Розмитість зору.
- Itching: Покраснення шкіри, свербіж.
- Irritability: Роздратування.
- Delayed healing: Повільне загоєння ран.
- Partial paresis: Часткова пареза (втрата частини м'язової сили).
- Muscle stiffness: Затвердіння м'язів.
- Alopecia: Випадіння волосся.
- Obesity: Ожиріння.
- Diagnosis: Діагноз діабету (1 - позитивний, 0 - негативний).

Code Book: Early Stage Diabetes Risk Prediction Dataset

Data Set Information:

Набір даних містить медичні записи пацієнтів та їх демографічну інформацію. Записи включають кілька клінічних ознак і цільову змінну, яка вказує, чи є у пацієнта ризик розвитку діабету на ранній стадії чи ні. Набір даних отримано зі сховища машинного навчання UCI.

Attribute Information:

- Age: age in years (numeric)
- Sex: Male or Female (binary)
- Polyuria: Increased urination (binary)
- Polydipsia: Increased thirst (binary)
- Sudden weight loss: rapid weight loss (binary)
- Weakness: weakness or fatigue (binary)
- Polyphagia: increased hunger (binary)
- Genital thrush: yeast infection (binary)
- Visual blurring: blurry vision (binary)
- Itching: itching or pruritus (binary)
- Irritability: irritability or mood swings (binary)
- Delayed healing: delayed wound healing (binary)
- Partial paresis: partial paralysis (binary)
- Muscle stiffness: stiffness in muscles (binary)
- Alopecia: hair loss (binary)
- Obesity: BMI > 30 (binary)
- Class: indicates if the patient is at risk of developing early stage diabetes or not (binary)

Таблиця **patient**:

Поле	Тип даних	Ключ
patient_id	Ціле число	Primary key
full_name	Рядок	-
gender	Рядок	-
birthday	Дата	-

Таблиця **doctor**:

Поле	Тип даних	Ключ
doctor_id	Ціле число	Primary key
full_name	Рядок	-

Таблиця **analysis**:

Поле	Тип даних	Ключ
analysis_id	Ціле число	Primary key
polyuria	Булево значення	-
polydipsia	Булево значення	-
sudden_weight_loss	Булево значення	-
partial_paresis	Булево значення	-
polyphagia	Булево значення	-
forecast_id	Ціле число	Foreign key (forecast)

Таблиця **forecast**:

Поле	Тип даних	Ключ
forecast_id	Ціле число	Primary key
forecast	Булево значення	-

Таблиця **patient_analysis**:

Поле	Тип даних	Ключ
patient_id	Ціле число	Foreign key (patient)
doctor_id	Ціле число	Foreign key (doctor)
analysis_id	Ціле число	Foreign key (analysis)
date	Дата	-