НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Факультет інформатики та обчислювальної техніки

Кафедра обчислювальної техніки

«На правах рукопису»
УДК _____

До захисту допущено:
Завідувач кафедри
_____ Сергій СТІРЕНКО
«___»_____2023 р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-професійною програмою «Інженерія програмного забезпечення комп'ютерних систем» зі спеціальності 121 “Інженерія програмного забезпечення” на тему: «Автоматизована система для діагностування цукрового діабету із використанням машинного навчання»

Виконав (-ла):
студент (-ка) ІІ курсу, групи ІМ-21мп
Кришталь Дмитро Вікторович _____

Керівник:
доц., к.т.н., доц.
Долголенко Олександр Миколайович

_____

Консультант з нормоконтролю:
проф. каф. ОТ, д.т.н., професор
Жабін Валерій Іванович

_____

Рецензент:
доц, к.т.н., доц
В’ячеслав Чмельов Орійович _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2023 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет *Інформатики та обчислювальної техніки*
(повна назва)

Кафедра *Обчислювальної техніки*
(повна назва)

Рівень вищої освіти – другий (магістерський)

Спеціальність *121. Інженерія Програмного Забезпечення*
(код і назва)

Освітньо-професійна програма *Комп'ютерні системи та мережі*
(код і назва)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Сергій СТІРЕНКО
(підпис)
«____» _____ 2023 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Кришталь Дмитро Вікторовичу
(прізвище, ім'я, по батькові)

1. Тема дисертації *Автоматизована система для діагностування цукрового діабету із використанням машинного навчання*

Науковий керівник дисертації *доц., к.т.н., доц. Долголенко Олександр Миколайович*
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «07» листопада 2023 р. № 5168-с

2. Строк подання студентом дисертації _____

3. Об'єкт дослідження *цукровий діабет*

4. Предмет дослідження *нейронні мережі, що використовуються для бінарної класифікації*

5. Перелік завдань, які потрібно розробити: *огляд літературних джерел, дослідження хвороби цукрового діабету, створення моделі машинного навчання*

6. Консультанти розділів дисертації:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Нормконтроль	Жабін В. І.		
1.	Долголенко О. М.		
2.	Долголенко О. М.		
3.	Долголенко О. М.		
4.	Долголенко О. М.		

7. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Строк виконання етапів дисертації	Примітка
1.	*Затвердження теми дослідження. Визначення предмету дослідження*	*02.09.2023*	
2.	*Дослідження літератури*	*16.09.2023*	
3.	*Аналіз існуючих підходів*	*30.09.2023*	
4.	*Створення моделі нейронної мережі*	*25.10.2023*	
5.	*Оформлення документації пояснювальної записки*	*05.11.2023*	
6.	*Попередній захист та проходження нормативного контролю*	*25.11.2023*	
7.	*Подання ДП рецензенту*	*28.11.2023*	
8.	*Захист*	*16.01.2024*	

Студент ____Дмитро КРИШТАЛЬ____ _____
(підпис)

Науковий керівник дисертації ___Олександр ДОЛГОНЕНКО___ _____
(підпис)

РЕФЕРАТ

на магістерську дисертацію

виконану на тему: «Автоматизована система для діагностування цукрового діабету із використанням машинного навчання»

Робота складається із вступу та чотирьох розділів. Загальний обсяг роботи: 67 аркушів основного тексту, 34 ілюстрацій, 17 таблиць. При підготовці використовувалася література з 41 різного джерела.

Актуальність: У сучасному світі є величезна кількість захворювань, які сильно поширюються по всьому земному шару. До таких можна віднести цукровий діабет, який є дуже серйозною хворобою, яка сильно впливає на стан кровоносних судин та й загалом на роботу внутрішніх органів. Максимально швидке його діагностування та призначення лікування може суттєво зменшити шкоду організму людини.

Машинне навчання набуває особливого значення у сучасних технологіях. Його можна використовувати в безлічі різних сфер, наприклад, у медицині. Воно ідеально підходить для діагностування таких хвороб як цукровий діабет, наявність якої можна визначити за допомогою певної множини характеристик здоров’я пацієнту.

Мета і завдання дослідження: метою магістерської дисертації є реалізація алгоритмів машинного навчання, за допомогою яких можна визначити наявність цукрового діабету у людини, порівняння їх роботи між собою.

Завдання досліджень:

- Аналіз хвороби цукрового діабету
- Аналіз поняття машинного навчання: парадигми та типи нейронних мереж.
- Пошук та аналіз датасетів
- Створення алгоритмів машинного навчання.
- Порівняння їх роботи

Об’єкт дослідження: нейронні мережі.

Предмет дослідження: нейронні мережі для бінарної класифікації.

Методи дослідження: для досягнення поставлених в магістерській роботі задач використано нейронні мережі, що використовують керований(навчання з учителем) тип навчання.

Особистий внесок здобувача. Магістерське дослідження є самостійно виконаною роботою, в якій відображено особистий авторський підхід та особисто отримані теоретичні та прикладні результати, що відносяться до вирішення задачі діагностування цукрового діабету з використання алгоритмів машинного навчання.

Практична цінність. Отримані результати можуть використовуватися у майбутніх дослідженнях за напрямками:

- Полегшення діагностики цукрового діабету.
- Автоматизація процесу діагностування.
- Вдосконалення методів медичної діагностики.
- Застосування у медичній освіті.

Ключові слова: Машинне навчання, штучний інтелект, цукровий діабет, автоматизована система, бінарна класифікація.

Пояснювальна записка

до магістерської дисертації

на тему: «Автоматизована система для діагностування цукрового діабету із використанням машинного навчання»

Київ – 2023

ЗМІСТ

ВСТУП

Останні декілька століть пов’язані зі стрімким рівнем розвитку технологій, значним збільшенням кількості підприємств, появою нових видів транспорту, тощо. Проте, разом з численними досягненнями приходить свідомість про серйозні проблеми в глобальному масштабі, такі як забруднення довкілля, зміна клімату та загрози здоров'ю населення. Але, на щастя, з розвитком раніше зазначених елементів, також і відбувся дуже сильний розвиток у сфері медицини, що дозволило не лише просто залишити тривалість життя на тому ж рівні, а й сильно збільшити її, не зважаючи на сучасні проблеми з екологією.

Спостереження за впливом забруднення навколишнього середовища на здоров'я показує зростання випадків захворювань, включаючи і цукровий діабет. Наявність ефективних систем діагностики та контролю за цією хворобою стає надзвичайно важливою, а машинне навчання відкриває нові можливості для покращення медичних практик та допомагає зменшити вплив забруднень на здоров'я.

Здоров’я населення планети стоїть перед викликом, який несе значний глобальний і соціальний вплив - поширення цукрового діабету. Ця недуга, що впливає на мільйони людей по всьому світу, стає дедалі більш поширеною і потребує ретельного моніторингу та діагностики. Щорічно діагноз "цукровий діабет" ставиться мільйонам людей, і це робить актуальним завдання розробки та вдосконалення автоматизованих систем для діагностики цієї захворюваності. У цьому контексті, використання сучасних методів машинного навчання стає важливим інструментом для створення точних та надійних систем діагностики цукрового діабету.

Цільове завдання цієї магістерської дисертації полягає у розробці та оцінці автоматизованої системи для діагностики цукрового діабету, яка використовує методи машинного навчання на основі клінічних та біохімічних даних пацієнтів. Цей дослідження спрямоване на покращення точності та ефективності діагностики цукрового діабету, що відкриває можливість для раннього виявлення та лікування цієї хвороби.

Цукровий діабет має серйозні наслідки для здоров'я пацієнтів, і раннє виявлення є вирішальним для успішного контролю за хворобою та запобігання її ускладненням. Автоматизована система для діагностики цукрового діабету, заснована на аналізі клінічних та біохімічних показників, може виокремити ризик цієї хвороби на ранній стадії та сприяти покращенню діагностичної точності.

У цій магістерській дисертації буде розглянуто ключові аспекти розробки та імплементації автоматизованої системи для діагностики цукрового діабету, включаючи збір клінічних та біохімічних даних, обробку та аналіз цих даних за допомогою методів машинного навчання, а також валідацію та оцінку точності системи. Результати цього дослідження можуть внести значний внесок у покращення діагностики цукрового діабету та сприяти зменшенню впливу цієї хвороби на здоров'я людей.

Ця магістерська дисертація висвітлить сучасний стан досліджень у сфері діагностики цукрового діабету, методи та інструменти машинного навчання, а також подробиці розробки автоматизованої системи та її можливі переваги для клінічної практики.

РОЗДІЛ 1
АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ. ПОНЯТТЯ ЦУКРОВОГО ДІАБЕТУ. ДІАГНОСТИКА

1.1 Вступ до проблеми.

Цукровий діабет - це хронічне захворювання, яке стало однією з найсерйозніших медико-соціальних проблем сучасності. Порушення обміну глюкози в організмі, притаманне цьому захворюванню, сильно впливає на функціонування внутрішніх органів, призводить до ускладнень і зниження якості життя пацієнтів. Така ситуація робить діабет однією з найважливіших глобальних загроз для здоров'я людини. Дана хвороба часто є причиною інших захворювань, які призводять навіть до смерті, таких як, хронічна ниркова недостатність, інфаркт міокарда або інсульт.

Швидке поширення цукрового діабету серед населення планети є надзвичайно тривожним явищем. За останні десятиліття кількість хворих на цю хворобу зросла вдвічі, і, за даними Всесвітньої організації охорони здоров'я, вже близько 422 мільйонів людей у світі страждають від цукрового діабету. Крім того, це значення надалі продовжує швидко зростати, що ставить під загрозу як глобальне здоров'я, так і економічну стійкість суспільства.

Причини швидкого поширення цукрового діабету серед населення:

- Зміни у способі життя: Зміни в харчуванні, збільшення вживання висококалорійних продуктів, недостатня фізична активність і стрес можуть призвести до збільшення ризику розвитку цукрового діабету.
- Ожиріння: Ожиріння є одним із ключових факторів ризику для діабету 2 типу.
- Збільшення тривалості життя: Загальна тривалість життя зросла, і багато людей досягли віку, коли ризик розвитку цукрового діабету збільшується. Як наслідок, збільшується кількість хворих.
- Збільшення діагностики: Зокрема, покращилася діагностика цукрового діабету, що дозволяє виявляти більше випадків хвороби.

Важливим аспектом у боротьбі з хворобою є як рання діагностика, так і лікування та підтримка пацієнтів. Рання діагностика діабету значно полегшує життя пацієнтів і запобігає розвитку серйозних ускладнень, забезпечуючи раннє лікування та контроль захворювання. Однак багато пацієнтів дізнаються про свою хворобу занадто пізно і вже мають серйозні наслідки на своє здоров'я

У цьому контексті розробка та впровадження автоматизованих систем діагностики цукрового діабету набуває особливого значення. Методи машинного навчання на основі клінічних і біохімічних даних можуть значно підвищити точність і раннє виявлення цього захворювання. Такі системи можуть стати потужним інструментом для медичних працівників.

Метою цього дослідження є розробка та оцінка автоматизованої системи діагностики діабету з використанням методів машинного навчання на основі клінічних та біохімічних даних. Система покликана підвищити точність діагностики, допомогти виявити ризик розвитку діабету на ранній стадії, покращити якість життя пацієнтів та зменшити соціально-економічні наслідки діабету.

1.2 Поняття та класифікація цукрового діабету.

1.2.1 Визначення цукрового діабету та його типу.

Цукровий діабет є хронічним медичним станом, який характеризується порушенням обміну глюкози в організмі. Основним симптомом цього недугу є підвищений рівень цукру (глюкози) в крові, що може виникнути через недостатню продукцію інсуліну, реакцію організму на цей гормон або комбінацію обох факторів.

На сьогодні розрізняють декілька типів цукрового діабету. До основних відносять наступні:

- Цукровий діабет 1-го типу (ЦД1): Цей тип хвороби є аутоімунним захворюванням, де імунна система організму атакує та руйнує клітини підшлункової залози, які виробляють інсулін. Пацієнти з

ЦД1 повинні щодня вводити інсулін, оскільки їхні органи не виробляють цей гормон.

- Цукровий діабет 2-го типу (ЦД2): Цей тип є більш поширеним і зазвичай пов'язаним з інсулінорезистентністю, коли клітини організму не відповідають на інсулін. Початково пацієнти з ЦД2 можуть лікуватися дієтою та фізичною активністю, але в подальшому може знадобитися призначення пероральних антидіабетичних препаратів або інсуліну.

1.2.2 Патогенез діабету та основні фактори ризику.

Діабет розвивається в результаті дії різних факторів, включаючи генетичну схильність і зовнішні чинники. Патогенез захворювання пов'язаний з порушенням функції підшлункової залози та інсулінорезистентними клітинами. Перше призводить до дефіциту інсуліну, а друге не дозволяє клітинам використовувати інсулін для забезпечення нормального метаболізму глюкози.

Основні фактори розвитку цукрового діабету:

- **Спадковість:** Якщо у вашій родині є випадки цукрового діабету, зростає ймовірність його розвитку.
- **Зайва вага або ожиріння:** Великий об'єм жиру в організмі може сприяти розвитку інсулінорезистентності.
- **Фізична неактивність:** Недостатній рівень фізичної активності може збільшити ризик розвитку ЦД2.
- **Вік:** Ризик цукрового діабету зростає зі старіння.
- **Синдром полікістозних яєчників (PCOS):** Жінки з PCOS мають підвищений ризик розвитку ЦД2.
- **Гестаційний діабет:** Цукровий діабет, розвинутий під час вагітності, може збільшити ризик подальшого розвитку хвороби.

1.2.3 Класифікація цукрового діабету за міжнародними стандартами

Для систематизації та класифікації цукрового діабету за міжнародними стандартами використовується диференціація на основі основних характеристик та механізмів розвитку хвороби. За класифікацією Всесвітньої організації

охорони здоров'я (ВОЗ) та Американською асоціацією діабету (ADA), цукровий діабет поділяється на такі основні типи:

- Цукровий діабет 1-го типу (ЦД1)
- Цукровий діабет 2-го типу (ЦД2)
- Гестаційний діабет (розвивається під час вагітності)
- Інші специфічні типи діабету (у тому числі генетичні або індуцовані ліками)
- Порушення глікемії на голодний шлунок
- Змішаний тип цукрового діабету

Ця класифікація допомагає враховувати особливості та механізми розвитку кожного типу цукрового діабету, що має важливе значення для діагностики та лікування.

1.3. Симптоми та ускладнення цукрового діабету.

1.3.1 Основні клінічні прояви цукрового діабету.

Клінічні прояви діабету різноманітні і залежать від типу та стадії діабету. Однак існують загальні симптоми, які легко розпізнати як ознаки діабету:

Основні симптоми діабету:

- **Полідипсія (підвищена спрага):** Цей симптом виникає через те, що високий рівень глюкози в крові спричиняє зневоднення і підвищену спрагу. Пацієнт постійно відчуває спрагу і потребує великої кількості рідини.
- **Поліурія (посилене виділення сечі):** порушення метаболізму глюкози призводить до підвищеного виділення сечі та частих походів до туалету, навіть вночі. Це суттєво впливає на якість сну пацієнта.
- **Поліфагія (підвищений апетит):** Порушення метаболізму глюкози може призвести до відчуття голоду навіть після прийому їжі. Пацієнти можуть відчувати постійну тягу до їжі та надмірне споживання їжі, що може призвести до збільшення ваги.

- **Втрата ваги:** Незважаючи на підвищений апетит, хворі на діабет часто втрачають вагу, оскільки не можуть використовувати глюкозу як джерело енергії. Втрата ваги особливо помітна у пацієнтів з діабетом 1 типу.
- **Втома і слабкість:** Нестача внутрішньоклітинного цукру призводить до загального відчуття слабкості та втоми. Пацієнти стають менш активними та мають менше енергії.
- **Нечіткість зору:** порушення метаболізму цукру впливає на функцію очей. Це може проявлятися у вигляді нечіткості зору та порушені з фокусуванням. Якщо рівень цукру в крові не контролюється протягом тривалого часу, це може призвести до серйозних порушень зору, включаючи діабетичну ретинопатію, яка може призвести до сліпоти.
- **Виразки та пухлини, які потребують багато часу для загоєння:** діабет впливає на здатність організму загоювати рани. Можуть розвиватися виразки та пухлини, які не загоюються швидко, особливо на ногах. Це може призвести до інфекцій та ускладнень.

1.3.2 Хронічні ускладнення цукрового діабету та їх вплив на якість життя пацієнті.

Наслідки на здоров’я людини можуть бути різноманітні. Найбільш розповсюдженими є наступні:

- **Серцево-судинні ускладнення.** Діабет підвищує ризик серцево-судинних захворювань, таких як інфаркт міокарда, інсульт, гіпертонія та атеросклероз. Ці ускладнення можуть призвести до серйозних пошкоджень серця та кровоносних судин.
- **Нефропатія (ураження нирок).** Діабет може пошкодити нирки і призвести до хронічної ниркової недостатності. Може знадобитися гемодіаліз або трансплантація нирки.
- **Ретинопатія (ураження сітківки).** Якщо діабетичну ретинопатію виявити та лікувати пізно, вона може призвести до погіршення зору

та сліпоти. Для виявлення та лікування цього ускладнення важливо регулярно проходити офтальмологічні обстеження.

- **Нейропатія (ураження нервової системи).** Діабетична нейропатія спричиняє ураження нервової системи, що призводить до втрати чутливості, болю, оніміння та паралічу. Це може вплинути на рухову активність і здатність вільно пересуватися.
- **Інші ускладнення:** діабет вражає кровоносні судини, шкіру та кістки. Наприклад, поганий кровообіг може призвести до появи виразок і пухлин на шкірі, особливо на стопах, що може спричинити інфекцію та ампутацію.

1.4. Методи та підходи до діагностики цукрового діабету.

1.4.1 Традиційні методи діагностики діабету.

Традиційні методи діагностики діабету включають низку лабораторних та інструментальних методів:

- **Вимірювання рівня глюкози в крові**. Вимірювання рівня глюкози в крові є важливим способом діагностики діабету. Це робиться шляхом уколу пальця голкою або шляхом забору венозної крові. Зазвичай рівень глюкози в крові вимірюють натщесерце та після прийому їжі.
- **Глікований гемоглобін (HbA1c).** Цей тест вимірює концентрацію глікованого гемоглобіну в крові і відображає середній рівень глюкози в крові за попередні два-три місяці. HbA1c є важливим показником кількості глюкози, що поглинається еритроцитами.
- **Тест на толерантність до глюкози (ТТГ).** Цей тест допомагає визначити, як організм реагує на глюкозу. Пацієнти приймають глюкозу, а через дві години вимірюють рівень глюкози в крові; якщо рівень глюкози підвищений через дві години, у них може бути діабет.
- **Аналіз сечі на глюкозу.** Вимірювання рівня глюкози в сечі - ще один спосіб діагностики діабету.

1.4.2 Неінвазивні методи діагностики діабету.

Розвиток сучасного медичного обладнання та технологій дозволив діагностувати діабет неінвазивними методами:

- **Системи безперервного моніторингу глюкози (CGM).** Системи, які вимірюють рівень глюкози в крові в режимі реального часу. Мікросенсори імплантуються під шкіру, а дані передаються на монітор або смартфон пацієнта. CGM надає детальну інформацію про зміни рівня глюкози протягом дня.
- **Неінвазивні глюкометри на основі інфрачервоного або комбінаційного розсіювання**. Ці пристрої вимірюють рівень глюкози без забору крові. Вони використовуються для безконтактного вимірювання рівня глюкози в слині, твердих тканинах та інших рідинах організму.
- **Мобільні додатки для моніторингу рівня глюкози.** Додатки для смартфонів та інших мобільних пристроїв не тільки контролюють рівень глюкози в крові, але й надають інші корисні функції, такі як щоденники харчування та активності.

1.4.3 Аналіз новітніх досягнень машинного навчання для діагностики діабету.

Машинне навчання (МН) стає все більш важливим інструментом у діагностиці діабету. Алгоритми машинного навчання здатні аналізувати великі обсяги даних і виявляти закономірності, які залишилися б непоміченими при спостереженні людини.

Машинне навчання використовується для:

- **Прогнозування ризику розвитку діабету.** Моделі машинного навчання можуть визначити ризик розвитку діабету у людини на основі історичних даних, таких як вік, стать, генетика та рівень активності.
- **Діагностика на основі зображень.** Машинне навчання аналізує рентгенівські знімки, КТ та інші зображення, щоб виявити ознаки діабету, такі як аномалії кровоносних судин і підшкірно-жирової клітковини.

- **Покращення лікування та контролю рівня глюкози.** Машинне навчання може допомогти покращити дозування інсуліну та рекомендації щодо харчування на основі рівня глюкози пацієнта та інших даних.
- **Прогнозування ускладнень**. Машинне навчання можна використовувати для виявлення можливих ускладнень діабету на ранній стадії, що дозволить лікарям вчасно вжити заходів для запобігання ускладнень.

Аналіз останніх розробок у використанні машинного навчання для діагностики діабету є важливою частиною дослідження, що допомагає поліпшити способи діагностики та лікування захворювання.

1.5 Обґрунтування використання машинного навчання для діагностики діабету.

1.5.1 Аргументи для використання машинного навчання при діагностування діабету.

Існує кілька важливих аргументів на користь використання машинного навчання для діагностики діабету:

Автоматизація і точність: машинне навчання може швидко і точно аналізувати великі обсяги даних і виявляти взаємозв'язки і закономірності, які важко виявити вручну. Це підвищує точність діагностики і дозволяє лікарям виявляти закономірності, які інакше могли б залишитися непоміченими.

Персоналізація та індивідуалізоване лікування: машинне навчання може створити індивідуальний підхід до діагностики та лікування діабету на основі особистих характеристик кожного пацієнта. Це оптимізує лікування та покращує контроль над хворобою.

Аналіз багатьох факторів: машинне навчання може одночасно враховувати багато факторів, що впливають на розвиток діабету, включаючи генетичні, екологічні, соціальні та фактори способу життя. Це дає більш повну картину стану пацієнта.

Раннє виявлення ризиків та ускладнень. Машинне навчання дозволяє розробляти моделі для раннього виявлення ризиків та ускладнень діабету. Це дає змогу вжити превентивних заходів і знизити ризики.

Постійне вдосконалення: машинне навчання може вчитися на великих обсягах даних і постійно вдосконалювати моделі. Це важливо для адаптації до нових даних і змін у характеристиках пацієнтів.

1.5.2 Огляд попередніх досліджень з використанням методів машинного навчання для діагностики.

Попередні дослідження з використанням методів машинного навчання для діагностики діабету показують великий потенціал цих підходів. Основні висновки цих досліджень полягають у наступному:

Розробка моделей прогнозування ризику. Машинне навчання дозволяє створювати моделі, які можуть з високою точністю прогнозувати ризик розвитку діабету на основі ряду клінічних і демографічних даних.

Персоналізоване лікування та контроль рівня глюкози: машинне навчання генерує індивідуальні рекомендації щодо лікування та контролю рівня глюкози для окремих пацієнтів.

Аналіз зображень і біомаркерів: машинне навчання використовується для аналізу зображень кровоносних судин, сітківки ока та інших біомаркерів, що вказують на діабет.

1.5.3 Переваги та обмеження машинного навчання в охороні здоров'я.

Перевагами машинного навчання в охороні здоров'я є

Здатність аналізувати великі обсяги даних: машинне навчання може аналізувати і враховувати великі обсяги клінічних і лабораторних даних, важливих для діагностики і прогнозування.

Персоналізація підходу до пацієнтів: машинне навчання може створити індивідуальний підхід до діагностики та лікування для кожного пацієнта.

Раннє виявлення ризиків та ускладнень: машинне навчання може допомогти виявити ризик розвитку діабету та його ускладнень на ранній стадії.

Однак існують певні обмеження:

Потреба у великих обсягах даних: машинне навчання потребує великих обсягів високоякісних даних для побудови ефективних моделей.

Конфіденційність та етичні питання: Збір та обробка медичних даних пов'язані з питаннями конфіденційності та етики, які необхідно ретельно враховувати.

Потреба в експертному контролі: машинне навчання не може замінити клінічний досвід лікарів і потребує моніторингу та перевірки результатів.

З огляду на ці переваги та обмеження, використання машинного навчання в діагностиці діабету є перспективним напрямком для досліджень і медичних розробок.

Висновок до розділу 1

У даному розділі обговорюється важливість діабету та методів його діагностики в сучасному світі. Це пов'язано зі стрімким зростанням поширеності діабету в усьому світі, що призвело до необхідності вдосконалення методів діагностики та лікування діабету.

У цьому розділі визначено основні терміни та поняття, пов'язані з діабетом, а також представлено класифікацію діабету відповідно до міжнародних стандартів. Також детально описані симптоми діабету та хронічні ускладнення, які впливають на якість життя пацієнтів.

Особливу увагу приділено способам діагностики діабету. Розглянуто традиційні методи, такі як аналіз рівня глюкози в крові, а також неінвазивні методи діагностики з використанням новітнього медичного обладнання та технологій. Крім того, в цій главі проаналізовано останні досягнення у використанні машинного навчання для діагностики діабету.

Загалом, у першому розділі створюється важливий контекст для подальших досліджень, визначаючи цілі та завдання, які будуть розглянуті в наступних розділах цієї дисертації.

РОЗДІЛ 2
ОГЛЯД ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

2.1 Поняття нейронної мережі.

Нейронна мережа - це метод штучного інтелекту, який навчає комп'ютери обробляти дані таким чином, що надихається роботою людського мозку. Це тип машинного навчання, який називається глибоким навчанням, і використовує взаємопов'язані вузли або нейрони в шаровій структурі, схожій на людський мозок. Вона створює адаптивну систему, яку комп'ютери використовують для навчання на власних помилках і постійного вдосконалення. Отже, штучні нейронні мережі намагаються вирішувати складні завдання, такі як підсумовування документів, розпізнавання лиць, прогнозування різних подій, з більшою точністю.

Нейронні мережі можуть допомогти комп'ютерам приймати розумні рішення з обмеженою участю людини. Це через те, що вони можуть навчитися та моделювати взаємозв'язки між вхідними та вихідними даними, які є нелінійними та складними.

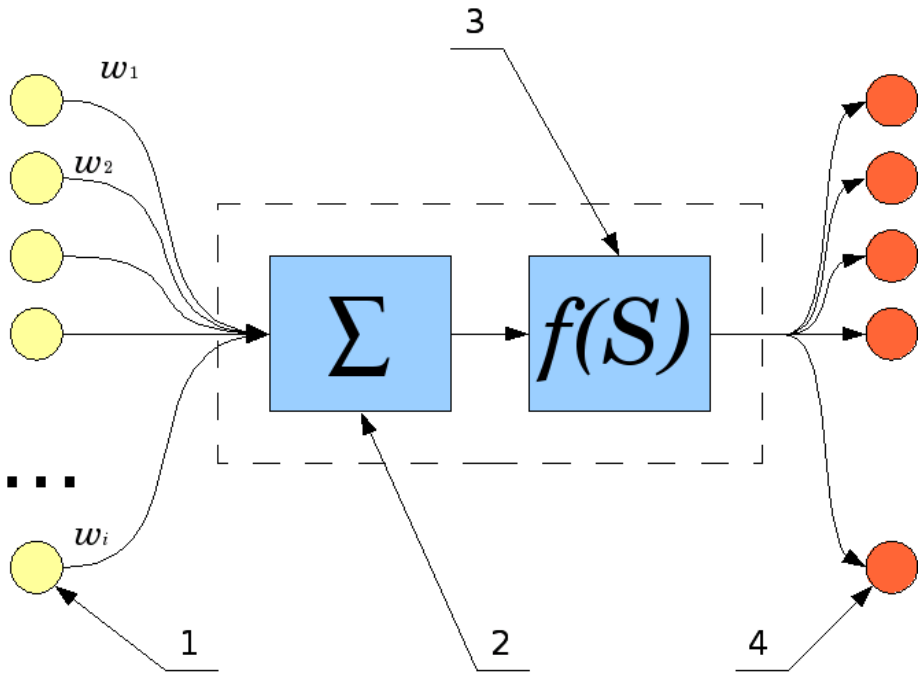

Рис 2.1. Модель штучного нейрону

2.2 Сфери, в яких використовуються нейронні мережі.

1. Медична: діагностування різних захворювань на основі певних даних
2. Таргетований маркетинг з використанням фільтрації соціальних мереж та аналізу поведінкових даних.
3. Фінансові прогнози на основі обробки історичних даних фінансових інструментів.
4. Прогнозування навантаження на електромережу та попиту на енергію.
5. Контроль якості та процесів.
6. Ідентифікація хімічних сполук.

2.3 Як працюють нейронні мережі.

Архітектура нейронних мереж натхнення людським мозком. Нейрони, які є клітинами людського мозку, утворюють складну, високо зв'язану мережу і передають один одному електричні сигнали, щоб допомогти людям обробляти інформацію. Аналогічно, штучна нейронна мережа складається з штучних нейронів, які спільно вирішують завдання. Штучні нейрони - це програмні модулі, які називають вузлами, і штучні нейронні мережі - це програмні системи або алгоритми, які, у своїй основі, використовують обчислювальні системи для вирішення математичних обчислень.

2.4 Проста архітектура нейронної мережі.

У базовій нейронній мережі є взаємопов'язані штучні нейрони в трьох шарах:

2.4.1 Вхідний Шар.

Інформація зовнішнього світу входить у штучну нейронну мережу через вхідний шар. Вузли входу обробляють дані, аналізують або категоризують їх і передають до наступного шару.

2.4.2 Прихований Шар.

Приховані шари отримують вхідні дані від вхідного шару або інших прихованих шарів. У штучних нейронних мережах може бути велика кількість прихованих шарів. Кожен прихований шар аналізує вихідні дані з попереднього шару, подальше їх обробляє і передає до наступного шару.

2.4.3 Вихідний Шар.

Вихідний шар надає остаточний результат всієї обробки даних штучною нейронною мережею. В ньому може бути один або кілька вузлів. Наприклад, якщо у нас є завдання бінарної (так/ні) класифікації, вихідний шар матиме один вихідний вузол, який надасть результат як 1 або 0. Однак, якщо у нас є завдання класифікації на кілька класів, вихідний шар може містити більше одного вихідного вузла.

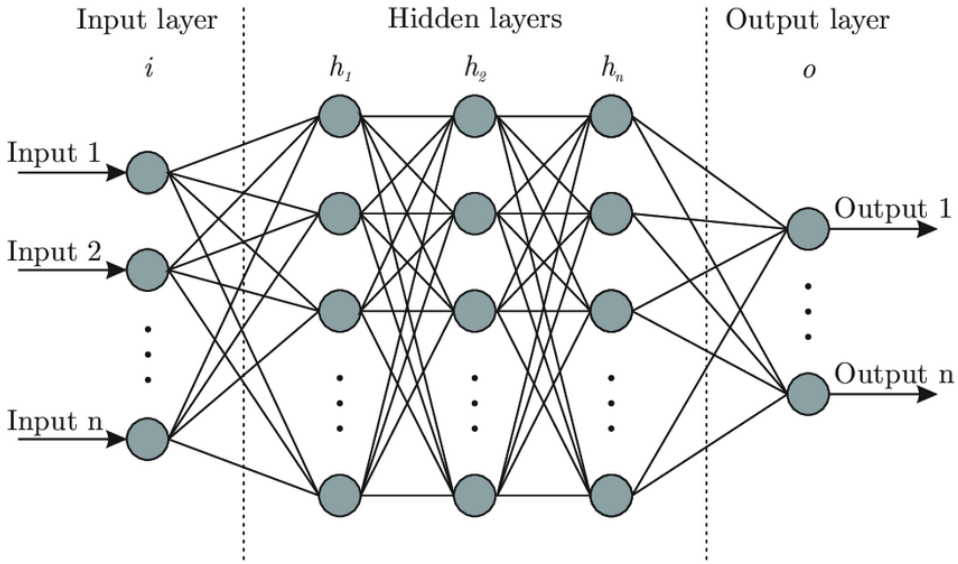

Рис 2.2. Нейронна мережа

2.5 Архітектура глибинної нейронної мережі.

Глибокі нейронні мережі, або мережі глибокого навчання, мають кілька прихованих шарів з мільйонами штучних нейронів, які пов'язані між собою. Вага, яка позначає зв'язки між одним вузлом і іншим, представляє собою числове

значення. Вага є позитивним числом, якщо один вузол стимулює інший, або негативним, якщо один вузол пригнічує інший. Вузли з вищими значеннями ваги мають більший вплив на інші вузли.

Теоретично глибокі нейронні мережі можуть відобразити будь-який тип вхідних даних на будь-який тип вихідних даних. Проте для їх навчання потрібно набагато більше часу порівняно з іншими методами машинного навчання. Їм потрібні мільйони прикладів тренувальних даних, а не сотні чи тисячі, які, можливо, знадобляться простішій мережі.

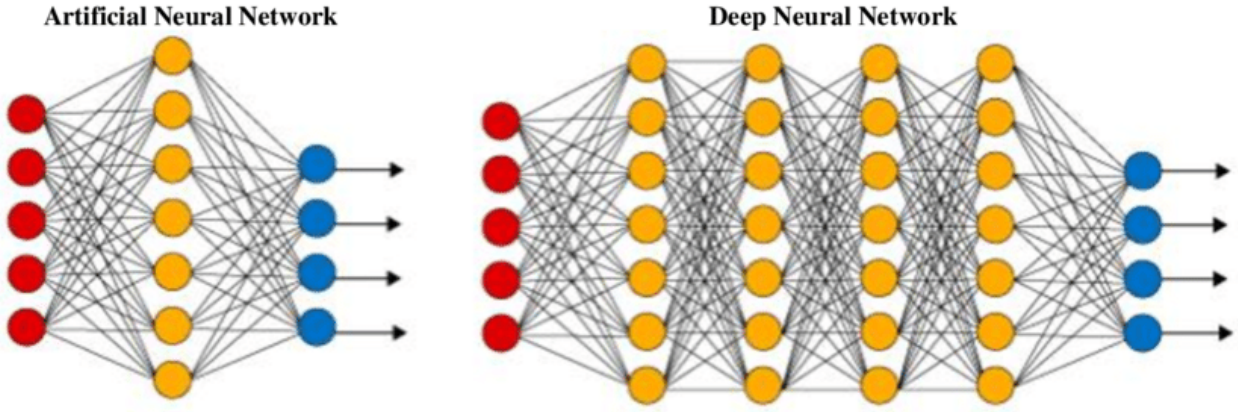

Рис 2.3. Порівняння штучної нейронної мережі з глибинною нейронною мережею

2.6 Типи нейронних мереж.

Штучні нейронні мережі можна розділити на категорії за тим, як дані пересуваються від вхідного вузла до вихідного вузла. Нижче наведені деякі приклади:

2.6.1 Нейронна мережа прямого поширення.

Нейронна мережа прямого поширення, або багатошарові перцептрони (MLP) - це тип штучної нейронної мережі, в якій інформація рухається в одному напрямку від входу до виходу через кілька шарів взаємопов'язаних нейронів. Ці мережі використовуються для різних завдань машинного навчання, таких як класифікація, регресія і розпізнавання образів. Ці мережі складаються з вхідного шару, прихованого шару і вихідного шару зі зваженими зв'язками між нейронами

і нелінійною функцією активації. Під час навчання вагові коефіцієнти налаштовуються для мінімізації помилки прогнозування. Нейронні мережі прямого поширення є основою глибокого навчання і застосовуються в таких сферах, як розпізнавання зображень і обробка природної мови.

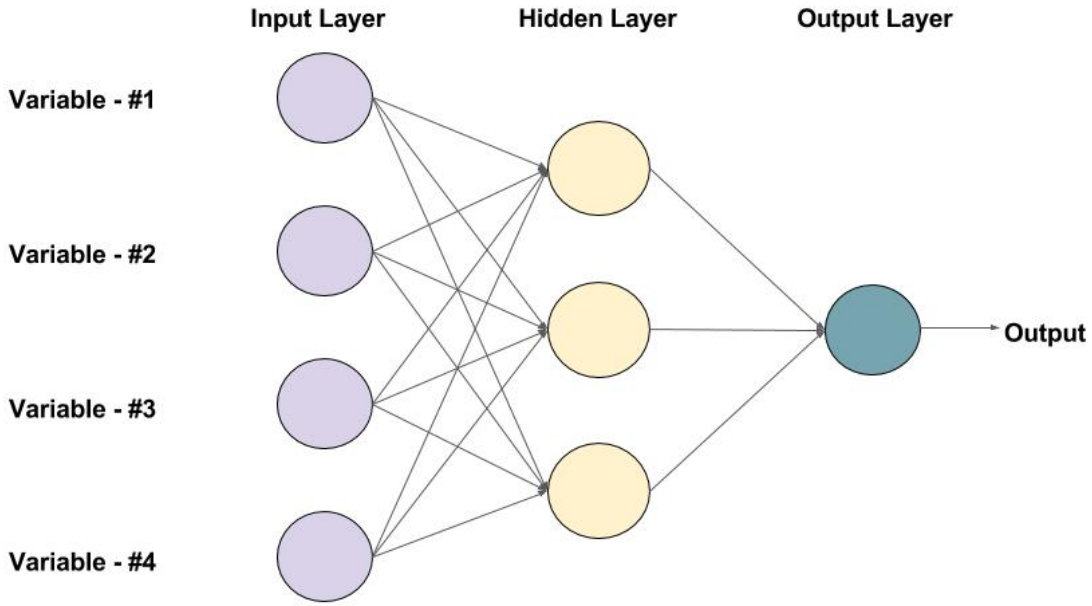

Рис 2.3. Нейронна мережа прямого поширення

2.6.2 Метод зворотного поширення помилки.

Штучні нейронні мережі безперервно навчаються і використовують коригувальний зворотний зв'язок для покращення прогнозного аналізу. Простіше кажучи, нейронна мережа може уявити, що дані проходять від вхідного вузла до вихідного вузла багатьма різними шляхами. Існує лише один правильний шлях від вхідного вузла до правильного вихідного вузла. Щоб знайти цей шлях, нейромережа використовує зворотній зв'язок:

Кожен вузол робить припущення про наступний вузол на шляху.

Він перевіряє, чи це припущення правильне. Вузли присвоюють більшу вагу шляхам, які ведуть до більш правильних припущень, і меншу вагу шляхам, які ведуть до неправильних припущень.

Для наступної точки даних вершина робить нове припущення, використовуючи шлях з більшою вагою, і повторює крок 1.

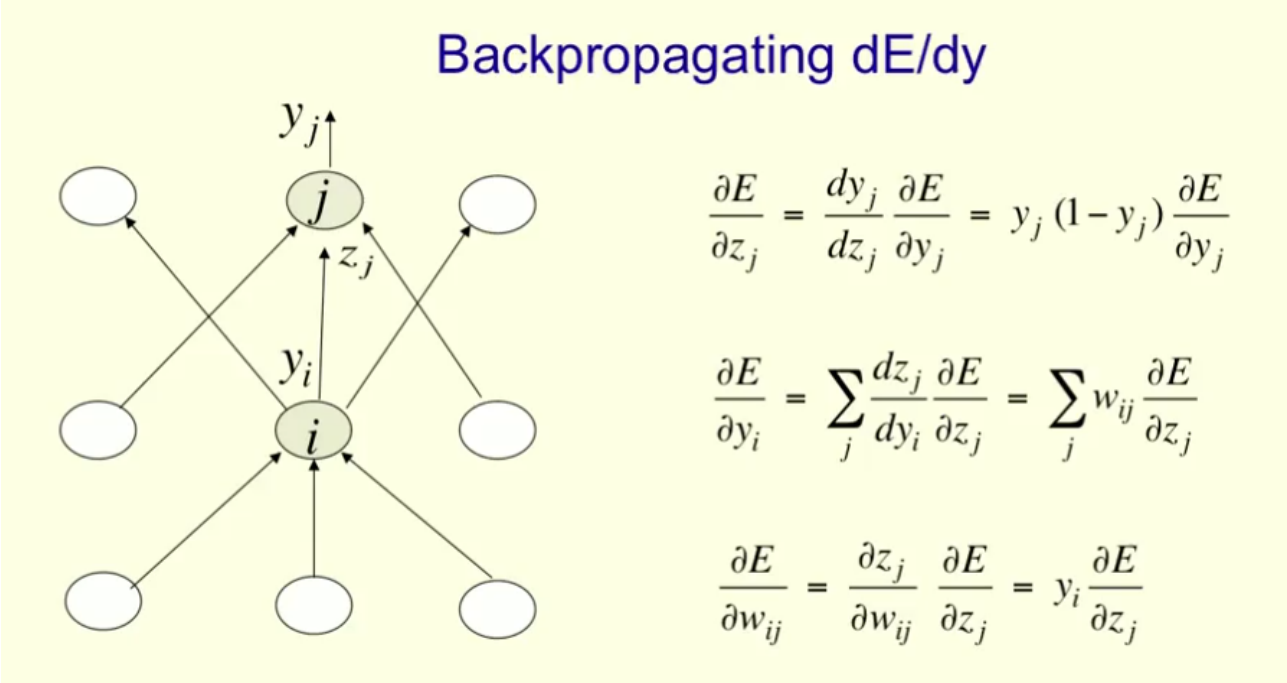

Рис 2.5. Метод зворотного поширення помилки

2.6.3 Згорткова нейронна мережа.

У згортковій нейронній мережі приховані шари виконують певні математичні функції, такі як узагальнення та фільтрація, відомі як згортка. Вони дуже корисні для класифікації зображень, оскільки можуть виокремлювати із зображень важливі ознаки, корисні для розпізнавання та класифікації зображень. Новий формат зображень легше обробляти, не втрачаючи важливих ознак, необхідних для точного прогнозування. Кожен прихований шар витягує і обробляє різні характеристики зображення, такі як контур, колір і глибина.

Шар згортки застосовує процес згортки до вхідних даних і передає результат наступному шару. Згортка імітує реакцію окремого нейрона на зоровий стимул.

Кожен згортковий нейрон обробляє дані лише з власного рецептивного поля.

Хоча повністю пов'язані нейронні мережі прямого поширення можуть використовуватися як для навчання ознак, так і для класифікації даних, застосовувати цю архітектуру до зображень непрактично. Оскільки розмір вхідних даних дуже великий і кожен піксель є цікавою змінною, зображення вимагають дуже великої кількості нейронів навіть у мілких (на відміну від глибоких) архітектурах. Наприклад, для (невеликого) зображення 100 x 100 у

повністю пов'язаному шарі є 10 000 ваг. Згорткова обробка вирішує цю проблему, оскільки зменшує кількість вільних параметрів і дозволяє створювати глибші мережі з меншою кількістю параметрів. Наприклад, 5 × 5 з'єднувальних областей з однаковими спільними вагами потребують лише 25 вільних параметрів, незалежно від розміру зображення. Таким чином, проблема градієнтного згасання та розриву при навчанні традиційних багатошарових нейронних мереж за допомогою зворотного поширення може бути вирішена.

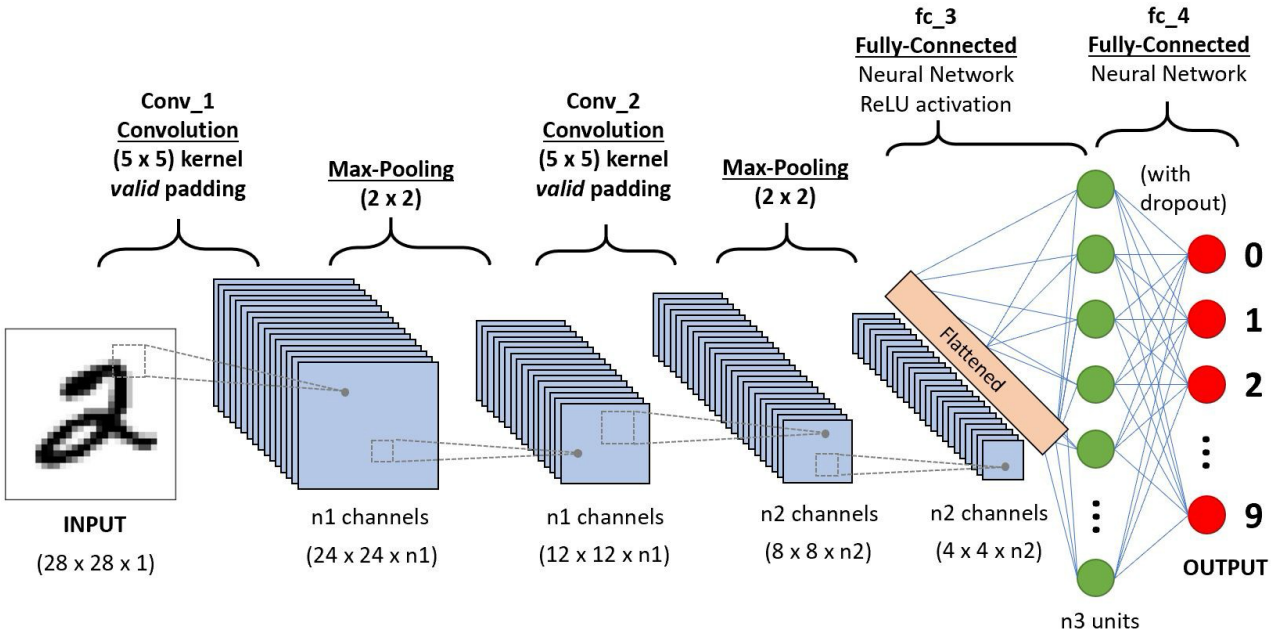

Рис 2.6. Приклад згорткової нейронної мережі

2.7 Парадигми машинного навчання.

Враховуючи показники ефективності, точності, надійності та пояснюваності, фахівці з обробки даних розглядають навчання з учителем (кероване навчання), навчання без учителя (некероване навчання), слабокероване навчання(напівкероване навчання) та навчання з підкріпленням для досягнення оптимальних результатів.

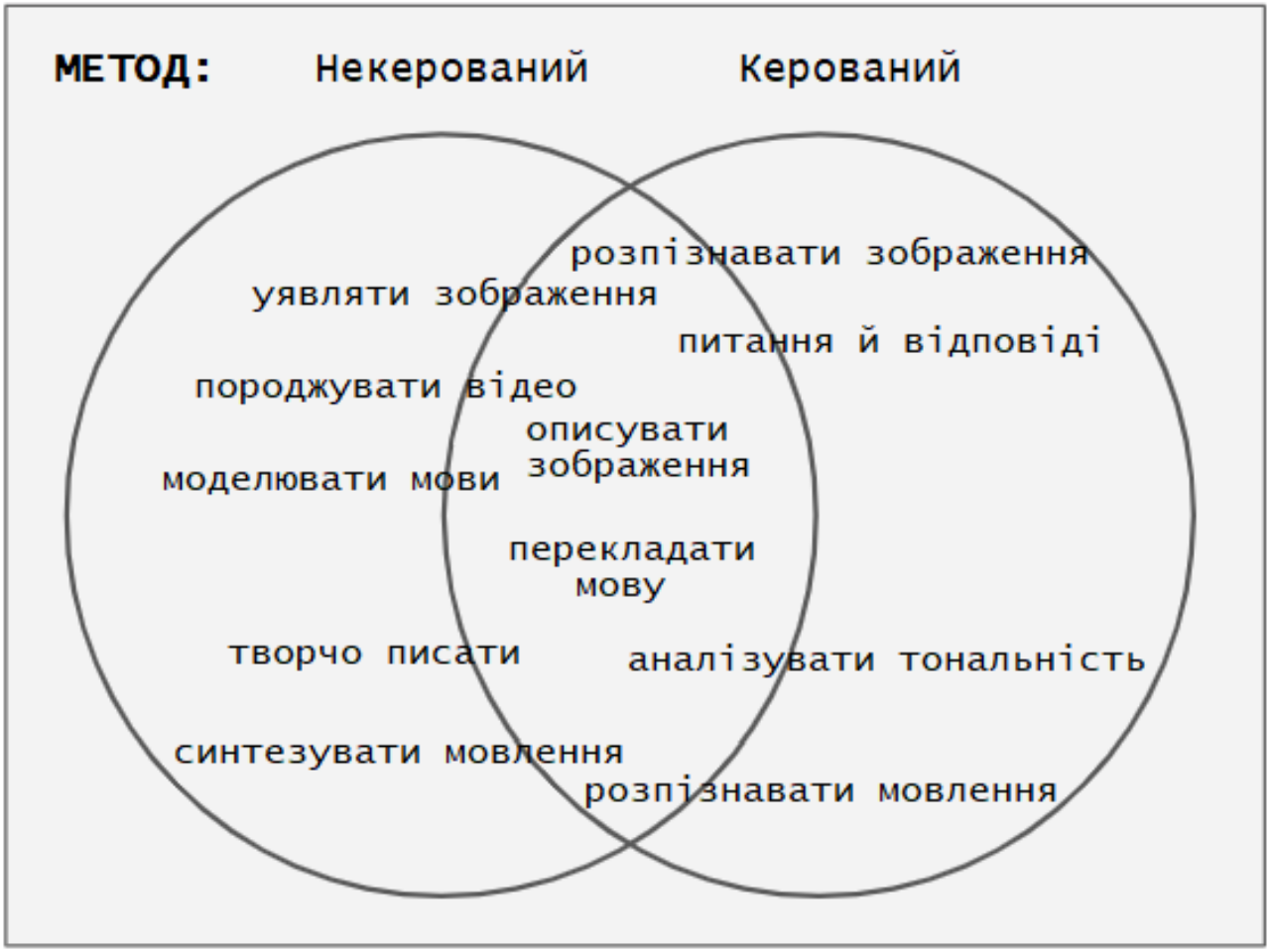

Рис 2.7. Відмінності керованого та некерованого навчання

2.7.1 Вибір парадигми машинного навчання.

Наука про дані починається експериментальним та ітеративним процесом з метою визначення найбільш правильного підходу у термінах ефективності, точності, надійності та пояснюваності. Типи машинного навчання стають корисними при розгляді різних переваг та недоліків певного класу алгоритмів для конкретної задачі, враховуючи походження даних. Існують практики машинного навчання, коли розробники комбінують різні типи машинного навчання та різноманітні алгоритми всередині цих типів для досягнення оптимальних результатів.

Вчені з обробки даних можуть аналізувати набір даних за допомогою технік ненавчання з метою отримання базового розуміння відносин всередині набору даних — наприклад, як продаж квартири корелює із його розташуванням у місті. Якщо цей зв'язок підтверджений, практики можуть використовувати техніки навчання з учителем із мітками, які описують розташування квартири у

місті. Техніки напівнавчання можуть автоматично розраховувати мітки розташування на полиці. Після впровадження моделі машинного навчання може бути використане навчання з підсиленням для уточнення передбачень моделі на основі фактичних продажів.

2.7.2 Навчання з учителем (кероване навчання).

Моделі навчання під наглядом мають справу з попередньо позначеними даними. Останні досягнення в галузі глибокого навчання з'явилися в результаті проекту Стенфордського університету в 2006 році. Недолік полягає в тому, що хтось або якийсь процес повинен наносити ці мітки. Це забирає багато часу і важко застосовувати мітки згодом. У деяких випадках мітки генеруються автоматично як частина автоматизованого процесу. Класифікація та регресія - найпоширеніші типи алгоритмів керованого навчання.

Алгоритми класифікації визначають категорії суб'єктів, об'єктів або подій, представлених у даних. Найпростіші алгоритми класифікації відповідають на бінарні питання, такі як "так/ні", "продаж/не продаж" і "кіт/не кіт". Більш складні алгоритми групують об'єкти в декілька категорій, наприклад, коти, собаки та миші. Найпоширеніші алгоритми класифікації включають дерева рішень та логістичну регресію.

Алгоритми регресії визначають взаємозв'язок між кількома змінними, представленими в наборі даних. Цей підхід корисний при аналізі того, як певна змінна, наприклад, продажі товару, пов'язана з такими змінними, як ціна, температура, день тижня, місце на полиці тощо. Поширені алгоритми регресії включають лінійну регресію, багатовимірну регресію, дерева рішень і регресію з використанням операцій додавання і зменшення (ласо).

Серед поширених застосувань - категоризація зображень об'єктів, прогнозування тенденцій продажів, класифікація кредитних заявок.

2.7.3 Навчання без учителя (некероване навчання).

У керованому навчанні, після того, як деякі характеристики вивчені, модель навчається співвідносити вхідні дані з вихідними, щоб набути здатності до узагальнення та правильної класифікації раніше не бачених вибірок даних. Однак вони можуть не знати, що є виходом, оскільки мають лише вхідні дані і не

можуть визначити вихідні мітки для кожної вхідної вибірки. Наприклад, уявімо, що ми працюємо в компанії, яка продає одяг, і у нас є дані про минулих клієнтів (сума покупки, вік і дата покупки). Наше завдання - знайти закономірності та взаємозв'язки між змінними і надати компанії корисну інформацію для розробки маркетингових стратегій і визначення, на яких клієнтах слід зосередитися, щоб максимізувати прибуток, або на яких сегментах клієнтів ми можемо зосередитися більше, щоб зростати на ринку. Ось для чого це потрібно.

У цьому випадку результатом не будуть мітки, оскільки наші потреби не можуть бути змодельовані таким чином. Натомість наш додаток повинен мати можливість групувати клієнтів відповідно до того, що є схожим або унікальним. Це групування базуватиметься на атрибутах, вивчених на етапі навчання. У цьому випадку підхід до навчання є неконтрольованим, оскільки немає керівника, який би надавав мітки для зіставлення вхідних і вихідних даних.

Моделі некерованого навчання автоматизують процес розпізнавання закономірностей, присутніх у наборі даних. Ці закономірності особливо корисні при дослідженні даних, щоб визначити, як найкраще вирішити проблеми науки про дані. Два найпоширеніші типи алгоритмів навчання без учителя - це кластеризація та зменшення розмірності.

Алгоритми кластеризації допомагають групувати схожі набори даних за різними критеріями. Експерти можуть розділити дані на різні групи і виявити закономірності всередині кожної групи.

Алгоритми зменшення розмірності шукають способи ефективного компактного представлення даних для певної проблеми.

Ці алгоритми включають в себе підходи відбору ознак і проекції. Відбір ознак допомагає визначити пріоритетність ознак, які є більш важливими для даної проблеми. Проекція ознак досліджує способи виявлення основних взаємозв'язків між кількома змінними, які можна кількісно виміряти в нові проміжні змінні, більш релевантні для проблеми, що розглядається.

Кластеризація та зменшення розмірності часто використовується для групування товарів на основі даних про продажі, зіставлення даних про продажі

з розташуванням товарів на полицях магазинів, класифікації клієнтів та ідентифікації ознак на зображеннях.

Наприклад, у нас є певна вибірка елементів, які ми хочемо згрупувати за допомогою некерованого навчання. Вони згруповані відповідно до ряду різних характеристик, які відрізняють їх один від одного.

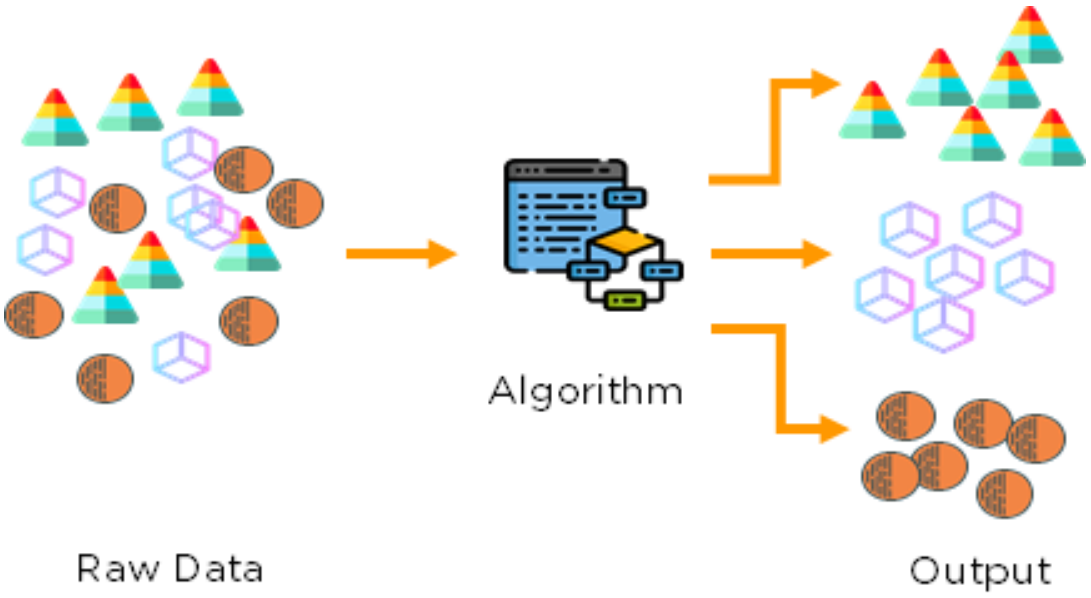

Рис 2.8. Приклад некерованого навчання

2.7.4 Порівняльна характеристика керованого та некерованого навчання.

Основна відмінність полягає в тому, що в керованому навчанні моделі навчаються передбачати результати на основі маркованого набору даних. Це означає, що він вже містить приклади правильних відповідей, які були ретельно відібрані людиною-наглядачем. З іншого боку, при неконтрольованому навчанні модель має справу з великою кількістю немаркованих даних і намагається зрозуміти їх без нагляду людини.

Таблиця 2.1.

Порівняння керованого навчання з некерованим

Властивості	Навчання з учителем	Навчання без учителя
Визначення	Тип машинного навчання під наглядом людини. Вхідні дані позначаються міткою (label), який	Різновид неконтрольованого машинного навчання. Машина самостійно

Таблиця 2.1 (Продовження)

Порівняння керованого навчання з некерованим

	показує машині бажаний результат.	виявляє закономірності в даних.
Вхідні дані	Дані з попередньо визначеними наглядачем(людиною) мітками(label)	Дані без попередньо визначених міток(label)
Використання даних	Модель задається вхідною змінною (X), вихідною змінною (Y) та алгоритмом навчання функції входу-виходу.	Модель використовує лише вхідну змінну (X) без відповідних вихідних даних та алгоритмом навчання функції входу-виходу.
Випадки, коли слід застосовувати	Коли людина знає, на які конкретні закономірності слід звертати увагу в множині раніше отриманих даних.	Коли, людина не впевнена, на що слід звертати увагу в множині даних
Використовується для	Задач, в яких слід класифікувати об’єкти за певними їхніми властивостями.	Задач, ціль яких полягає у кластеризації та асоціації
Точність кінцевих результатів	В результаті роботи алгоритму створеного за принципами навчання без учителя зазвичай спостерігаються більш точні результати, оскільки перелік відповідей, які може надати система	В результаті роботи алгоритму створеного за принципами навчання без учителя зазвичай спостерігаються менш точні результати.

Таблиця 2.1 (Продовження)

Порівняння керованого навчання з некерованим

	зазвичай обмежується двома(так/ні, собака/кішка)	

2.7.5 Напівкероване навчання (слабке керування).

Моделі напівкерованого навчання описують процес використання алгоритмів некерованого навчання для автоматичного створення міток для даних, які можуть бути використані в методах керованого навчання. Існує кілька підходів до застосування цих міток, зокрема

1. Методи кластеризації позначають дані мітками, подібними до міток, створених людиною.
2. Методи навчання, що самоорганізуються, навчають алгоритми розв'язувати кінематичні задачі, які правильно застосовують мітки.
3. Методи навчання на основі багатьох екземплярів знаходять способи генерувати мітки для колекцій екземплярів з певними властивостями.

2.7.6 Навчання з підкріпленням.

Моделі навчання з підкріпленням часто використовують для вдосконалення моделей після їх впровадження. Моделі навчання з підкріпленням також можна використовувати для інтерактивних процесів навчання, наприклад, для навчання алгоритму гри на основі зворотного зв'язку про окремі кроки або для прийняття рішення про те, хто виграє чи програє в раунді гри, наприклад, шахмати.

Основний підхід полягає у створенні набору дій, параметрів і кінцевих точок, які встановлюються методом проб і помилок. Алгоритм робить прогнози, ходи і рішення на кожному кроці. Результати порівнюються з результатами гри або реальним сценарієм. Надсилаються покарання та винагороди, а алгоритм з часом вдосконалюється.

Найпоширеніші алгоритми навчання з підкріпленням використовують різні нейронні мережі. Наприклад, у додатках для автономного водіння алгоритми можуть навчатися на основі того, як вони реагують на дані, записані з

транспортного засобу, або на синтетичні дані, що представляють те, що датчики транспортного засобу можуть бачити вночі.

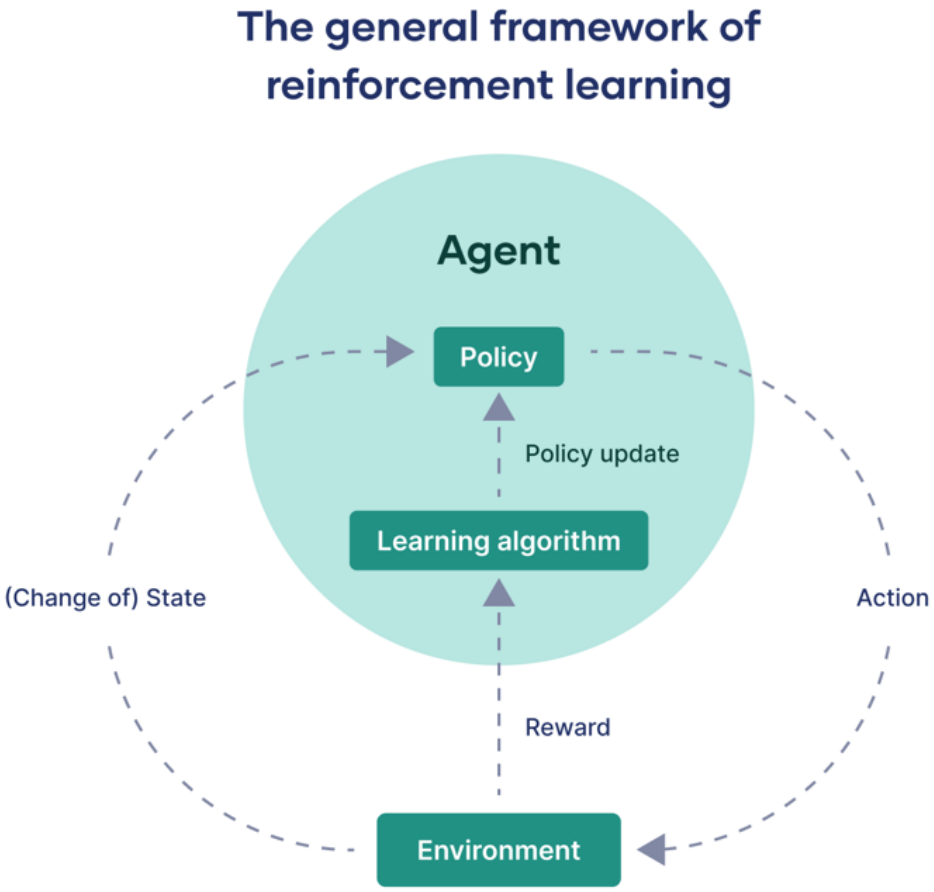

Рис 2.9. Приклад роботи алгоритму навчання з підкріпленням

2.8 Огляд технологій для розробки системи.

2.8.1 Мова програмування Python.

Python - це високорівнева мова програмування загального призначення, відома своєю простотою та читабельністю коду. Цей розділ містить детальний огляд мови програмування Python:

- **Простота і читабельність:** Python відома своїм чистим і зрозумілим синтаксисом. Мова розроблена таким чином, щоб полегшити розробникам читання та розуміння коду, що робить її ідеальною для початківців.
- **Інтерпретована мова:** Python є інтерпретованою мовою, що означає, що код може бути виконаний без необхідності його компіляції. Код

можна виконувати і налагоджувати по одному рядку або функції за раз.

- **Кросплатформеність:** Python підтримується багатьма операційними системами, включаючи Windows, macOS та різні дистрибутиви Linux. Ваш код на Python може бути відтворений на різних платформах без значних змін.
- **Універсальність:** Python підтримує різні парадигми програмування, включаючи процедурне, об'єктно-орієнтоване та функціональне програмування. Її можна використовувати для створення різних типів додатків.
- **Багата бібліотечна екосистема:** Python має бібліотеки та модулі для наукових обчислень (NumPy, SciPy), обробки даних (Pandas), візуалізації (Matplotlib, Seaborn), машинного навчання (scikit-learn, TensorFlow, PyTorch) та багатьох інших сфер, включаючи Існує низка бібліотек та модулів, які спрощують розробку.
- **Велика та активна спільнота:** Python має велику та активну спільноту розробників. Це означає, що легко знайти документацію, форуми та підтримку для вирішення проблем і запитів.
- **Веб-розробка:** Python використовується для створення веб-додатків, особливо з використанням таких фреймворків, як Django та Flask.
- **Наукові дослідження та обробка даних:** Python використовується для наукових досліджень та обробки даних за допомогою таких бібліотек, як NumPy, SciPy, Pandas та Matplotlib.
- **Машинне навчання та штучний інтелект:** Python має широкий спектр бібліотек для розробки моделей машинного навчання та інтелектуальних систем, включаючи TensorFlow, PyTorch та scikit-learn.
- **Мова скриптів:** Python добре підходить для написання скриптів, автоматизації завдань та маніпулювання файлами і текстами.

Python є однією з найпопулярніших мов програмування у світі завдяки своїй простоті, універсальності та багатій екосистемі бібліотек. Вона підходить для широкого спектру завдань, від освіти та досліджень до веб-розробки та машинного навчання.

2.8.2 Бібліотека машинного навчання TensorFlow.

TensorFlow - це бібліотека з відкритим вихідним кодом для чисельних обчислень і машинного навчання, розроблена компанією Google. Вона надає потужні інструменти для побудови та навчання різних моделей штучного інтелекту, включаючи нейронні мережі. Цей розділ містить детальний огляд бібліотеки TensorFlow:

Обчислення графів: Однією з ключових особливостей TensorFlow є використання обчислювальних графів для опису та виконання обчислень. Граф - це структура, де вузли представляють операції, а ребра - дані, що передаються між операціями. Це дозволяє TensorFlow оптимізувати обчислення і працювати на різних пристроях, включаючи CPU і GPU, ефективно використовуючи апаратне забезпечення.

Гнучкість і масштабованість: TensorFlow підтримує створення різних типів моделей, таких як нейронні мережі, рекурентні мережі та згорткові мережі.TensorFlow можна використовувати для різноманітних завдань, від класифікації зображень до обробки природної мови та рекомендаційних систем.TensorFlow можна використовувати для різноманітних завдань, від класифікації зображень до обробки природної мови та рекомендаційних систем.

Широка спільнота та підтримка: TensorFlow має велику і активну спільноту розробників, яка надає документацію, навчальні посібники, приклади і рішення для різних завдань машинного навчання. Це полегшує навчання та розробку з цією бібліотекою.

Інтеграція з Keras: TensorFlow забезпечує інтеграцію з високорівневою бібліотекою Keras, що спрощує створення і навчання нейронних мереж; моделі можна швидко будувати за допомогою Keras, а бібліотеку можна використовувати для широкого спектру завдань машинного навчання.

Модуль TensorFlow Lite: TensorFlow має модуль TensorFlow Lite для розгортання моделей на мобільних пристроях і вбудованих системах, що дозволяє створювати додатки ШІ і машинного навчання на різних платформах.

Візуалізація та налагодження: TensorFlow надає інструменти для візуалізації обчислювальних графіків, налагодження коду та моніторингу процесу навчання моделі.

Застосування: TensorFlow використовується в різних галузях, включаючи обробку зображень, обробку природної мови, розпізнавання мови, робототехніку і наукові дослідження.

Спрощення розробки моделей ШІ: Завдяки високорівневому API та підтримці глибокого навчання, TensorFlow спрощує процес побудови та навчання складних моделей ШІ.

TensorFlow є потужним інструментом для розробників, що працюють в області машинного навчання і штучного інтелекту: він надає інфраструктуру для побудови, навчання і розгортання моделей, маніпулювання числовими даними і виконання різних завдань в області інтелектуальних систем TensorFlow надає інфраструктуру для виконання різних завдань в області інтелектуальних систем, включаючи

2.8.3 Бібліотека Keras.

Keras - це високорівнева бібліотека для роботи з нейронними мережами, вбудована в TensorFlow. Цей розділ містить детальний огляд бібліотеки Keras:

Простота використання: Keras розроблена для простоти і зручності використання. Вона має чистий та інтуїтивно зрозумілий інтерфейс, що дозволяє розробникам швидко створювати, навчати та оцінювати нейронні мережі.

Загальна відкритість: Keras підтримує різні бекенди, включаючи TensorFlow, Theano і Microsoft Cognitive Toolkit (CNTK), що дозволяє розробникам вибрати найбільш підходящий бекенд для конкретного завдання.

Модульність: Бібліотека Keras побудована за принципом модульності, що дозволяє легко комбінувати шари та процеси для створення складних нейромережевих архітектур.

Підтримка згорткових та рекурентних мереж: Keras підтримує як згорткові, так і рекурентні нейронні мережі, що дозволяє розробникам створювати різноманітні моделі, включаючи обробку зображень, тексту та часових рядів.

Перенесення моделей: Оскільки Keras є абстракцією різних бекендів, моделі, створені в Keras, можна легко переносити між різними бекендами. Це робить бібліотеку корисною для розробки моделей, які потребують різних бекендів для різних завдань.

Документація та спільнота: Keras має велику спільноту розробників та детальну документацію, яка допомагає розробникам зрозуміти та використовувати бібліотеку для різних завдань машинного навчання та глибокого навчання.

Популярність: Keras стала дуже популярним інструментом в індустрії глибокого та машинного навчання. Його використовують для побудови та навчання нейронних мереж у багатьох сферах, включаючи обробку зображень, обробку природної мови, генерацію тексту та багато інших завдань.

Підтримка TensorFlow 2.0: З виходом TensorFlow 2.0, Keras став офіційним API верхнього рівня для цього бекенду, що робить інтеграцію Keras і TensorFlow ще простішою.

Бібліотека Keras - це потужний інструмент для розробників, які шукають простий та ефективний спосіб побудови та навчання нейронних мереж. Вона робить процес розробки моделей глибокого навчання звичним та інтуїтивно зрозумілим, дозволяючи зосередитися на суті поставленого завдання.

Висновок до розділу 2

На даному етапі роботи на магістерською дисертацією було проведено огляд поняття машинного навчання та інструментів для розробки, які будемо використовувати у наступному розділі.

Також було зроблено огляд різних типів нейронних мереж та порівняння декількох основних парадигм навчання, які можуть бути використані для вирішення нашої проблеми.

Можна дійти висновку, що найкраще підходить нейронна мережа, яка використовує навчання з учителем, а саме та модель, якій для навчання необхідні атрибути та мітки, що завчасно визначені та вказують на правильний результат (так/ні)

Було проведено огляд інструментів, якими необхідно оволодіти для створення моделі, а саме мова Python, бібліотеки Keras та TensorFlow. У наступному розділі буде проведено огляд датасетів та розробку моделі, яка найбільш якісно зможе використовувати дана інструменти.

РОЗДІЛ 3
НАВЧАННЯ МОДЕЛЕЙ ТА РЕЗУЛЬТАТИ РОБОТИ

3.1 Огляд вибраних датасетів.

У цьому розділі описано ключові аспекти набору даних, які використовуються для навчання та оцінки моделі. Вибір набору даних визначається його репрезентативністю, розміром та відповідністю поставленому завданню.

На сьогоднішній день у мережі Інтернет можна знайти безліч різноманітних датасетів, які можна використовувати для навчання моделей машинного навчання. Найкращим ресурсом вважається Kaggle, який дозволяє не тільки отримати доступ до ряду інструментів та ресурсів для дослідження, вивчення та практики у сферах аналізу даних та розробки моделей машинного навчання, але й надає чудову можливість до участі у конкурсах з обробки даних, де учасники з усього світу змагаються у вирішенні реальних завдань зі збору та аналізу даних.

Я обрав 2 датасети, які, на мою думку, найкраще підійдуть для навчання моделі.

Таблиця 3.1.

Характеристики датасетів

Набір даних	Кількість записів	Кількість характеристик
1	768	9(Включно з міткою, яка позначає фактичну наявність хвороби(так/ні))
2	100,000	9(Включно з міткою, яка позначає фактичну наявність хвороби(так/ні))

3.2 Детальний огляд першого датасету.

3.2.1 Опис атрибутів.

Цей датасет містить атрибути, які пов’язані зі здоров’ям жінок. Він включає в себе наступні атрибути:

1. **Pregnancies (кількість вагітностей)**
   - **Опис:** кількість вагітностей у жінки. Кількість вагітностей впливає на ризик розвитку діабету, зокрема гестаційного діабету. Гестаційний діабет спричинений недостатнім виробленням інсуліну або неефективним використанням інсуліну організмом під час вагітності. Цей стан зазвичай розвивається у другому триместрі вагітності і може покращитися після пологів, але він також може підвищити ризик розвитку діабету 2 типу в подальшому житті.

     Таким чином, частота вагітності може бути фактором ризику через кілька механізмів:

     1) Збільшення навантаження на підшлункову залозу: Кожна нова вагітність створює додаткове навантаження на підшлункову залозу, яка виділяє інсулін. Підвищене інсулінове навантаження призводить до виснаження бета-клітин підшлункової залози.

     2) Метаболічні зміни: Гормональні та метаболічні зміни відбуваються з кожною вагітністю. Ці зміни впливають на чутливість клітин до інсуліну і можуть визначати розвиток інсулінорезистентності.
   - **Тип даних:** ціле число.
2. **Glucose (рівень глюкози в крові)**
   - **Опис**: Рівень глюкози в крові показує кількість цукру (глюкози) в крові в певний момент часу. Глюкоза є основним джерелом енергії для клітин організму, і її регуляція в крові важлива для нормального функціонування організму.

Рівень глюкози в крові вимірюється в міліграмах на децилітр (мг/дл) або мілімолях на літр (ммоль/л). Нормальний рівень може змінюватися в залежності від ряду факторів, включаючи час доби і час, що минув з моменту останнього прийому їжі.

Зміни рівня глюкози в крові можуть вказувати на низку станів, включаючи діабет, переддіабетний стан, гіперглікемію (високий рівень глюкози) та гіпоглікемію (низький рівень глюкози). Медичні працівники можуть визначити, як високий і низький рівень глюкози в крові впливає на ваше здоров'я, і розробити стратегію лікування або контролю цих станів.

- **Тип даних:** число з рухомою комою.

3. **Blood pressure (Кров’яний тиск)**

- **Опис:** розрізняють два види кров’яного тиску. У нашому випадку використовують діастолічний тиск (нижній показник).

  Артеріальний тиск є важливим показником функціонування системи кровообігу і вказує на тиск, який чинить на стінки артерій кров, що циркулює в судинах. Цей тиск визначається двома числами, що вимірюються в міліметрах ртутного стовпчика (мм рт.ст.)

  Систолічний артеріальний тиск (більше число): відображає максимальний тиск крові в артеріях під час скорочення серця, коли кров викидається в судини.

  Діастолічний артеріальний тиск (нижнє число): відображає найнижчий тиск в артеріях, коли серце розслаблюється між серцевими скороченнями.

  Наприклад, "120/80 мм рт.ст." означає, що систолічний тиск становить 120 мм рт.ст., а діастолічний - 80 мм рт.ст.

  Вважається, що нормальний артеріальний тиск у дорослих становить близько 120/80 мм рт.ст. Зміни цих значень можуть вказувати на ряд станів, включаючи підвищений артеріальний тиск (гіпертонію), знижений артеріальний тиск (гіпотонію) або

інші проблеми з артеріальним тиском, які впливають на серцево-судинну систему і загальний стан здоров'я. Оцінка артеріального тиску є важливою частиною здоров'я серця і враховується при діагностиці та лікуванні серцево-судинних захворювань.

- **Тип даних:** ціле число

4. **Skin thickness (товщина шкіри)**

- **Опис:** товщина шкіри. Товщина шкіри зазвичай вимірюється в певних розмірах і може бути використана як одна з багатьох характеристик для визначення ризику діабету та прогнозу, але сама по собі не є прямим показником діабету.

  Існує кілька обґрунтувань для використання товщини шкіри в контексті дослідження діабету. Наприклад, товщина підшкірно-жирової клітковини може бути пов'язана з резистентністю до інсуліну - важливим фактором розвитку діабету 2 типу. Інсулінорезистентності означає, що клітини організму менш чутливі до інсуліну, що призводить до підвищення рівня глюкози в крові.

  Однак товщина шкіри сама по собі не є точним і надійним показником діабету, і її використання може бути обмеженим. Діагноз діабету зазвичай ставиться за допомогою спеціальних тестів, таких як рівень глюкози в крові та інші біомаркери.

  Тому, хоча товщина шкіри може бути включена в аналіз ризику і дослідження характеристик людей з високим ризиком діабету, вона не може бути єдиним критерієм для виявлення діабету.

- **Тип даних:** ціле число

5. **Insulin (рівень інсуліну в крові)**

- **Опис:** кількість інсуліну в крові. Інсулін - це гормон, що виділяється бета-клітинами підшлункової залози, органу, розташованого в задній частині шлунка, і є важливим для регуляції рівня цукру (глюкози) в крові. Основна роль інсуліну - регуляція обміну речовин, особливо вуглеводів.

Основними функціями інсуліну в організмі є

1) Зниження рівня глюкози в крові: інсулін допомагає клітинам засвоювати глюкозу з крові і знижує концентрацію глюкози в крові.
2) Зберігання глюкози у формі глікогену: інсулін сприяє утворенню глікогену - форми зберігання глюкози в печінці та м'язах.
3) Стимуляція синтезу білка: інсулін сприяє вивільненню внутрішньоклітинних амінокислот і полегшує синтез білка.
4) Накопичення жиру: гормони сприяють утворенню та накопиченню жиру в клітинах.
5) Пригнічення розщеплення глікогену та жирів: інсулін підтримує стабільний рівень енергії, пригнічуючи розщеплення глікогену та жирів.
6) Дисфункція бета-клітин і клітинна резистентність до інсуліну можуть призвести до метаболічних порушень, включаючи розвиток діабету 2 типу, при якому організм не може ефективно використовувати інсулін або не може виділяти достатню кількість інсуліну.

- **Тип даних:** ціле число

6. **BMI (індекс маси тіла):**
   - Опис: відношення ваги до зросту. Особи з найвизим показником ІМТ (середній 34,5 кг/м²) мали 11-кратно збільшений ризик розвитку діабету порівняно з учасниками з найнижчим ІМТ (середній 21,7 кг/м²). Група з найвищим ІМТ мала вищий ймовірність розвитку діабету порівняно з усіма іншими групами ІМТ, незалежно від генетичного ризику.
   - Тип даних: число з рухомою комою.
7. **DiabetesPedigreeFunction (Функція родинної вибірки діабету):**
   - **Опис:** число, що відображає відношення кількості хворих на діабет до загальної кількості людей у родині. Оскільки ймовірність набуття

хвороби цукрового діабети у період життя людиною дуже сильно залежить від генетики, слід враховувати такий показник, як кількість хворих на дану хворобу у родині.

- **Тип даних:** число з рухомою комою.

8. **Age (вік):**
   - **Опис:** вік жінки. Даний показник впливає на ризик розвитку діабету, особливо діабету 2 типу. Основними способами, якими вік впливає на цей ризик, є:
     1) Збільшення ризику з віком: Загальна тенденція полягає в тому, що ризик розвитку діабету зростає з віком. Статистика показує, що люди старше 45 років мають вищий ризик розвитку діабету 2 типу, і цей ризик зростає з часом.
     2) Метаболічні зміни: з віком можуть відбуватися зміни в обміні речовин, наприклад, зниження чутливості клітин організму до інсуліну. Це призводить до інсулінорезистентності, яка є важливим фактором розвитку діабету 2 типу.
     3) Зміни у складі тіла: З віком тіло жінки може стати більшим, а розподіл жирової тканини може змінитися. Надмірна вага та ожиріння є факторами ризику розвитку діабету.
     4) Зниження фізичної активності: з віком фізична активність знижується, що може погіршити чутливість до інсуліну і загальний стан здоров'я.
   - **Тип даних:** ціле число.
9. **Outcome (результат):**
   - **Опис:** результат, де 1 вказує на наявність діабету, а 0 - відсутність.
   - **Тип даних:** бінарний (1 або 0).

Цей набір даних використовується для дослідження факторів, що впливають на розвиток діабету у жінок, а також для побудови моделей машинного навчання для прогнозування наявності або відсутності захворювання. Збалансованість та різноманітність характеристик роблять цей набір даних цікавою темою для дослідження та аналізу.

3.2.2 Аналіз записів датасету.

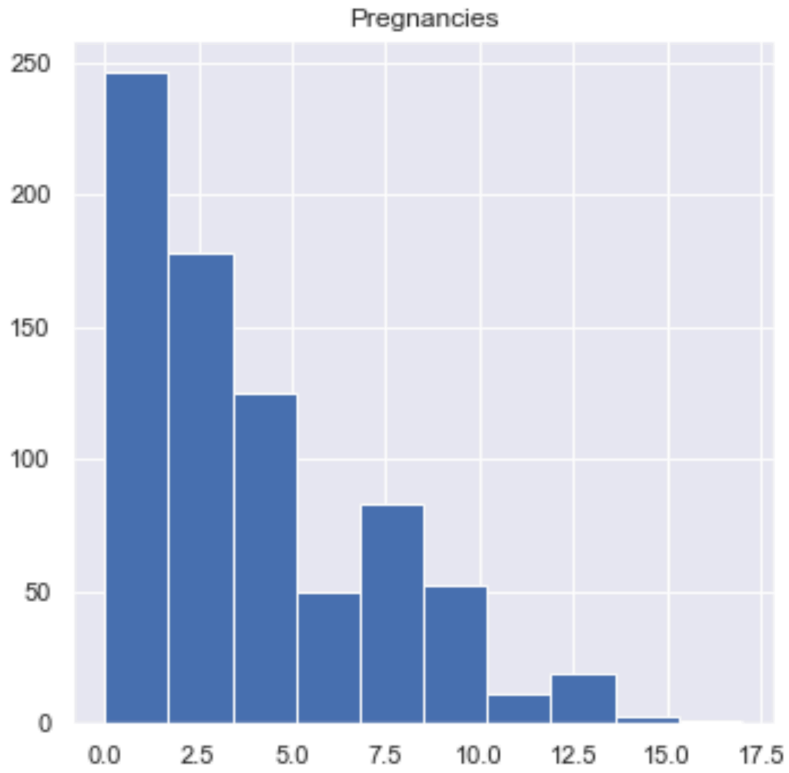

Рис 3.1. Діаграма розподілу за віковими групами у наборі 1.

З аналізу гістограм розподілу кількості вагітностей серед жінок можна зробити такі висновки:

1. Більшість жінок мають менше шести вагітностей: багато спостережень знаходяться в діапазоні від 0 до 6 вагітностей, що свідчить про те, що більшість жінок досліджуваної групи мають відносно невелику кількість вагітностей.
2. Кількість жінок з більш ніж шістьма вагітностями значно зменшується, коли перевищується інтервал у шість вагітностей. Це може свідчити про те, що в цьому датасеті було більше жінок з меншою кількістю вагітностей, що не є дивно, оскільки більшість країн має тенденцію до зменшування народжуваності.

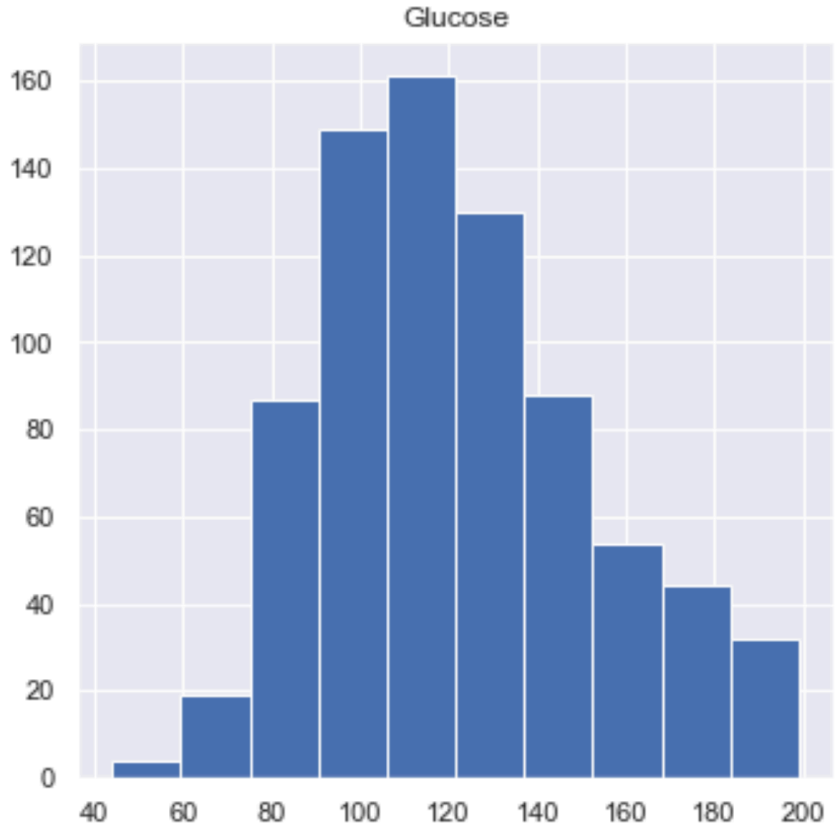

Рис 3.2. Діаграма розподілу за рівнем глюкози у наборі 1.

З аналізу гістограм розподілу кількості жінок за рівнем глюкози в крові можна зробити такі висновки:

1. Більшість жінок мають нормальний рівень глюкози в крові: відповідно до загального розподілу, більшість жінок мають рівень глюкози в крові в межах норми від 79,6 до 119,4. Кількість жінок у цих діапазонах (156 і 211 відповідно) є значною.
2. Частка жінок з низьким або високим рівнем глюкози невелика: дослідження показує, що дуже мало жінок мають рівень глюкози нижче 19,9 або вище 199. Такий розподіл свідчить про те, що екстремальні рівні глюкози в досліджуваній групі зустрічаються рідко.

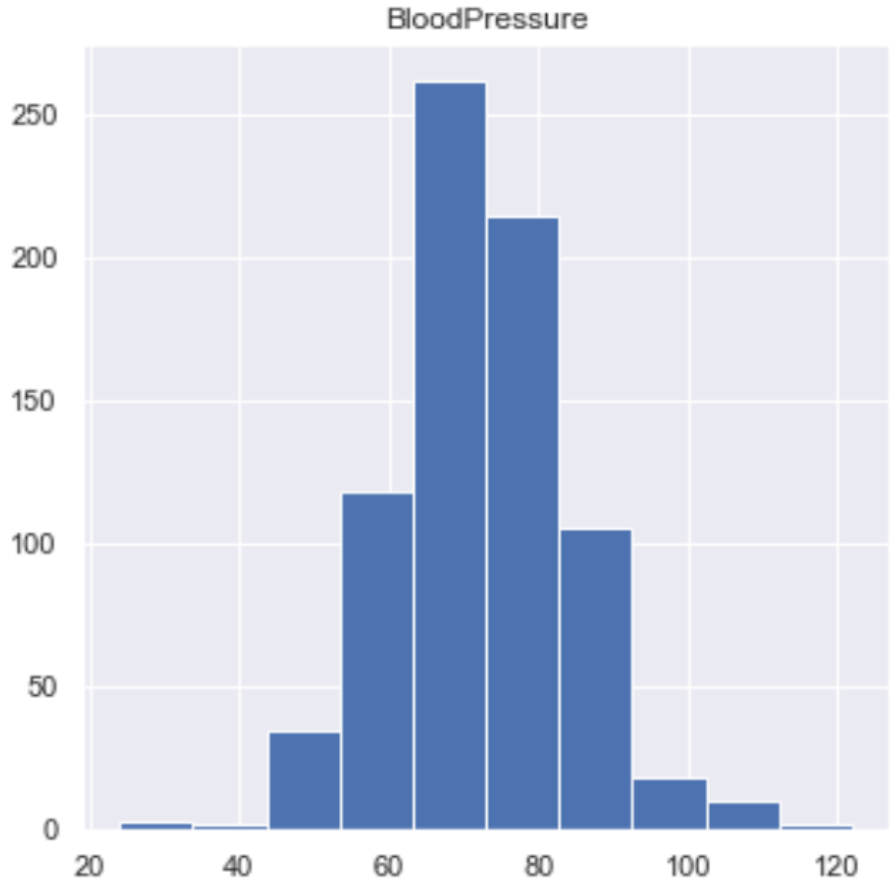

Рис 3.3. Діаграма розподілу за рівнем нижнього артеріального тиску у наборі 1

На основі аналізу гістограми розподілу кількості жінок за зниженням рівня артеріального тиску можна зробити такі висновки:

1. Більшість жінок мають нормальний рівень артеріального тиску: загальний розподіл показує, що більшість жінок мають артеріальний тиск у нормальному діапазоні (61-97,6). Цей діапазон є зоною нормального артеріального тиску, і кількість жінок у цьому діапазоні є значною (по 87-261 жінці).
2. Дуже мало жінок мають низький і високий артеріальний тиск: Діапазони артеріального тиску менше 24,4 і більше 109,8 є рідкісними серед досліджуваної популяції. Це свідчить про те, що екстремальна гіпотонія в досліджуваній популяції зустрічається рідко.

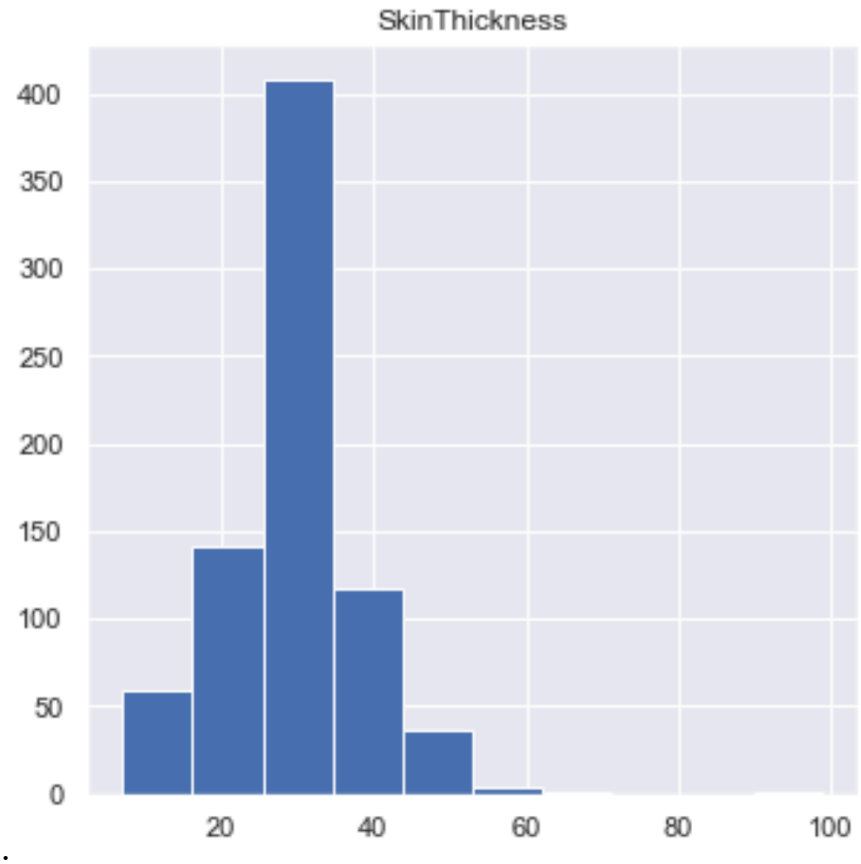

Рис 3.4. Діаграма розподілу за рівенем товщини шкіри у наборі 1.

З аналізу гістограм розподілу кількості жінок за товщиною шкіри можна зробити такі висновки

1. Товщина шкіри більшості жінок знаходиться в межах норми: Відповідно до загального розподілу, більшість жінок мають товщину шкіри в нормальному діапазоні від 9,9 до 49,5. Кількість жінок у цьому діапазоні дуже велика.
2. Дуже мало жінок мають дуже малу або дуже велику товщину шкіри: Діапазони товщини шкіри нижче 9,9 і вище 59,4 є невеликими в досліджуваній популяції. Це свідчить про те, що екстремальні значення товщини шкіри в досліджуваній популяції зустрічаються рідко.

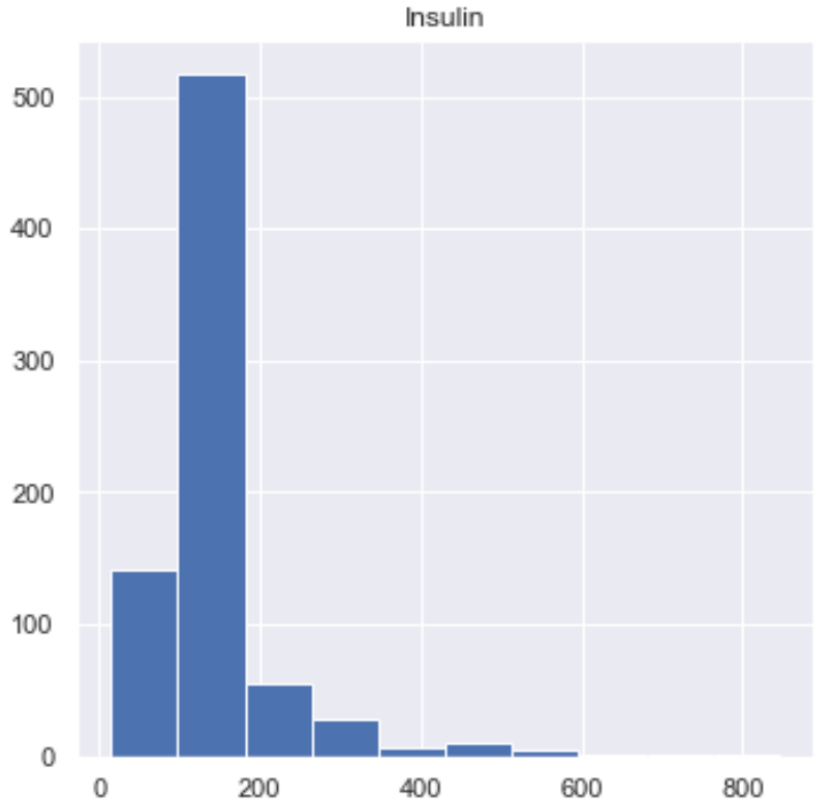

Рис 3.5. Діаграма розподілу за рівнем інсуліну у наборі 1.

На основі аналізу гістограми розподілу кількості жінок в залежності від рівня інсуліну можна зробити наступні висновки:

1. Більшість жінок мають низький рівень інсуліну: За загальним розподілом, більшість жінок мають низький рівень інсуліну, оскільки значущий пік кількості жінок спостерігається в діапазоні від 0 до 84.4. Це може вказувати на те, що у більшості учасниц дослідження рівень інсуліну знаходиться в нормі.
2. Менша кількість жінок із помірним рівнем інсуліну: Діапазон інсуліну від 84.6 до 169.2 має меншу кількість жінок порівняно з низьким рівнем. Це може вказувати на те, що ті, хто має помірний рівень інсуліну, представлені меншою кількістю

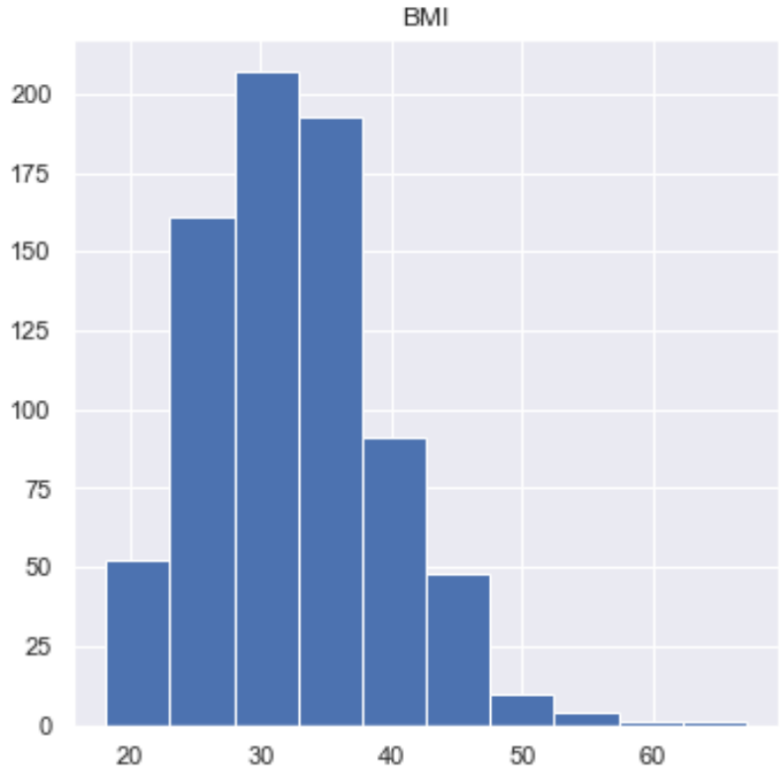

Рис 3.6. Діаграма розподілу за рівнем BMI у наборі 1

На основі аналізу гістограм розподілу кількості жінок за рівнями індексу маси тіла (ІМТ) можна зробити такі висновки:

1. Розподіл жінок з різними значеннями ІМТ: гістограми показують розподіл жінок з різними значеннями ІМТ. ІМТ більшості учасниць дослідження коливався від 20,1 до 33,5, що є діапазоном між нормальною та надмірною вагою за даними Всесвітньої організації охорони здоров'я (ВООЗ).
2. Піки в зоні нормальної та надмірної ваги: є піки в зоні нормальної (26,8-33,5) та надмірної ваги (20,1-26,8). Це може свідчити про те, що учасниці з такими значеннями ІМТ частіше зустрічаються в цьому дослідженні.
3. Небагато жінок з екстремальними значеннями ІМТ: Невелика кількість значень ІМТ в діапазонах нижче 20,1 і вище 40,2 свідчить про те, що в цьому дослідженні було мало учасниць з низькими і високими значеннями ІМТ.

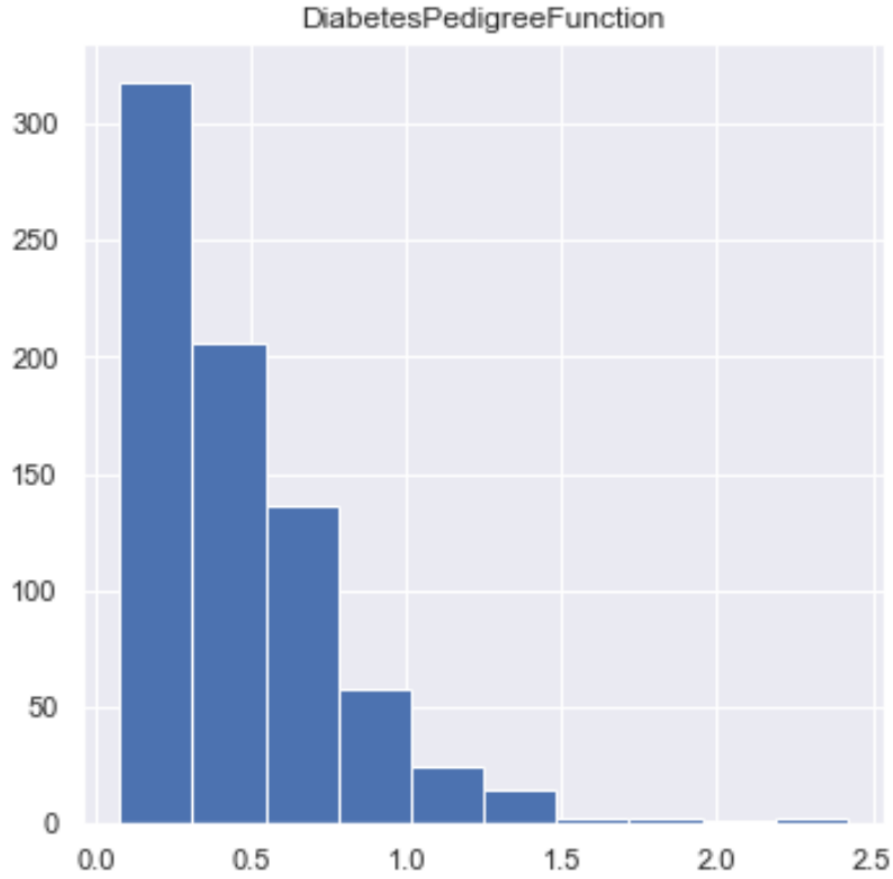

Рис 3.7. Діаграма родинної вибірки діабету у наборі 1.

На основі аналізу гістограм розподілу кількості жінок за рівнем DiabetesPedigreeFunction можна зробити такі висновки:

1. Загальний розподіл за рівнем DiabetesPedigreeFunction: Гістограми показують розподіл жінок у різних діапазонах значень DiabetesPedigreeFunction. Більшість учасниць дослідження мають значення DiabetesPedigreeFunction від 0 до 0,3.
2. Менше жінок мають високі значення DiabetesPedigreeFunction: Дуже мало жінок мають показник вище 0,55 (включно), що свідчить про те, що високі значення DiabetesPedigreeFunction не дуже поширені серед досліджуваної популяції.
3. Невелика частка жінок з дуже високими або дуже низькими значеннями функції: Частка жінок в інтервалах вище 1,25 і нижче 0,3 є невеликою, що свідчить про те, що кількість жінок з дуже високими або дуже низькими значеннями функції діабету за спадковістю є обмеженою.

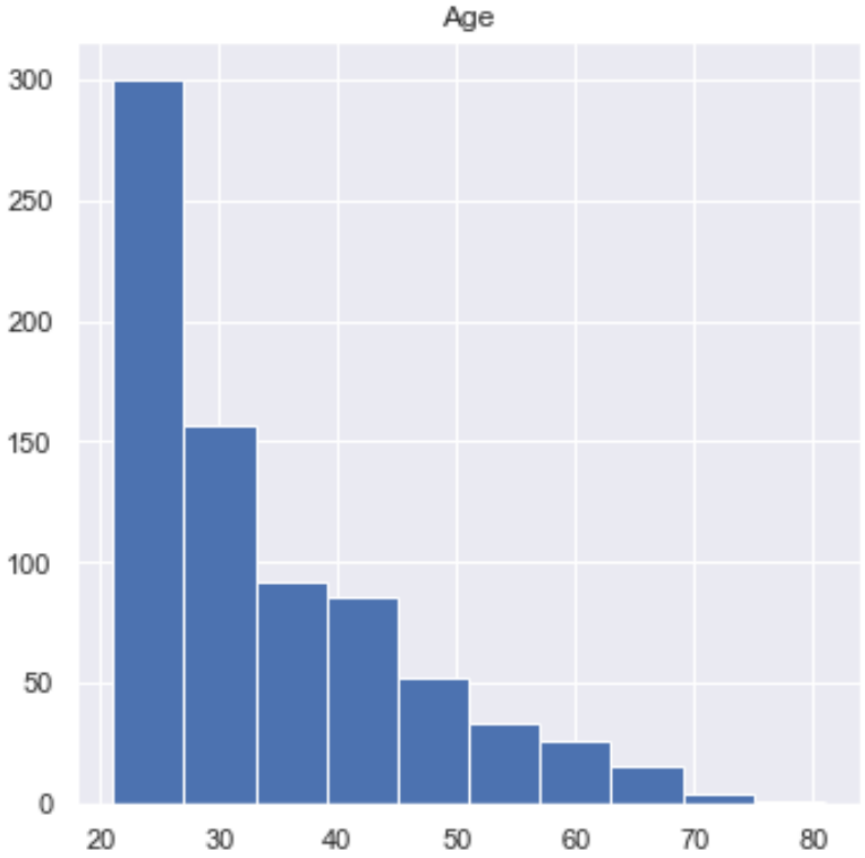

Рис 3.8. Діаграма розподілу за віком у наборі 1.

На основі гістограми за віком можна дійти висновку, що загалом жінки, які наявні у даному датасеті,

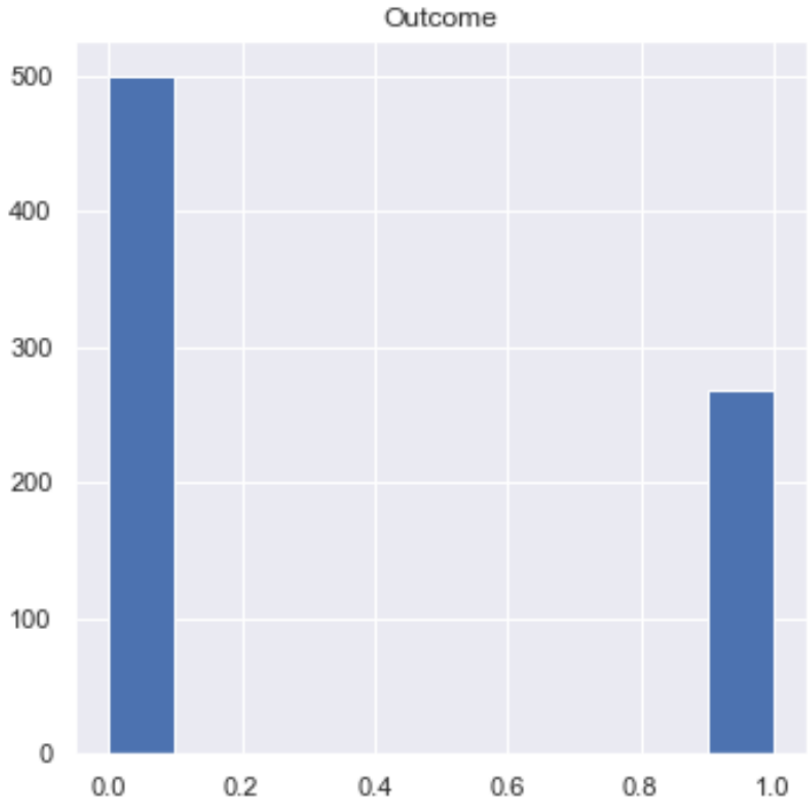

Рис 3.9. Діаграма розподілу за результатом у наборі 1.

3.2 Проектування моделі штучного інтелекту.

Після проведення огляду датасетів слід приступити до створення моделі штучного інтелекту. Для проектування моделі будемо використовувати сервіс Google Colab, який використовує Jupyter Notebooks. Їх можна використовувати для запуску та розробки коду на Python у хмарному середовищі; важливою перевагою Google Colab є вільний доступ до високопродуктивних графічних процесорів (GPU) та тензорних процесорів (TPU), які можна використовувати для прискорення виконання завдань штучного інтелекту (наприклад, навчання нейронних мереж).

3.2.1 Аналіз бібліотек необхідних для написання моделі

Для створення моделі спочатку слід визначитись з множиною бібліотек, які допоможуть у написанні даного проекту. Я обрав наступні:

1. Matpotlib
2. Seaborn
3. Pandas
4. Numpy
5. Missingno
6. Sklearn

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
import missingno as msno
from sklearn import metrics
```

Рис 3.10. Імпортування бібліотек

Таблиця 3.1.

Опис обраних бібліотек

Бібліотека	Опис	Використання
Matplotlib (import matplotlib.pyplot as plt)	Бібліотека для побудови графіків та візуалізації даних для Python. Вона надає широкі функціональні можливості для створення різних типів графіків, від лінійних до гістограм і контурних діаграм.	Візуалізація даних, дослідження залежностей та створення привабливих графіків.
Seaborn (import seaborn as sns):	Високорівнева бібліотека для візуалізації даних на основі Matplotlib. Вона додає рівні абстракції та надає стилізацію і високорівневі функції для швидкого створення стильних графіків.	Слід використовувати для створення привабливих графіків та візуалізацій даних.
Pandas (import pandas as pd):	Бібліотека для обробки та аналізу даних; вона надає структури даних, такі як DataFrame, і дозволяє легко працювати з табличними даними; за допомогою Pandas можна легко обробляти та аналізувати великі обсяги даних.	Використовується для завантаження, обробки та аналізу даних, особливо у вигляді табличних структур.

Таблиця 3.1. (Продовження)

Опис обраних бібліотек

NumPy (import numpy as np):	Бібліотека для обчислювальної математики, що використовується мовою Python. Вона надає високопродуктивні масиви та операції над ними, дозволяючи легко обробляти числові дані.	Він підтримує операції з масивами, векторами та матрицями і є дуже корисним для наукових обчислень та обробки даних.
missingno (import missingno as msno):	Бібліотека для візуалізації відсутніх значень у наборі даних. Дозволяє швидко оцінити кількість та місцезнаходження відсутніх даних у вигляді матриць та графіків.	Використовується на етапі аналізу даних для виявлення та візуалізації відсутніх значень.
sklearn (from sklearn import metrics)	Бібліотека машинного навчання для Python, що містить різні алгоритми для класифікації, регресії, кластеризації та інших задач машинного навчання. У цьому випадку імпортується її підмодуль metrics, який містить різні метрики для оцінки якості моделей машинного навчання.	Він використовується для оцінки продуктивності моделей машинного навчання для обчислення таких показників, як точність, повторюваність і метрики F1.

3.2.2 Підключення Google Drive до Google Colab та читання датасетів

Підключимо Google Drive до Google Colab за допомогою команди drive.mount(). Це дозволить отримати доступ до файлів, які зберігаються на ньому. У нашому випадку це файли датасетів.

```
from google.colab import drive
drive.mount('/content/drive')
```

Рис 3.11. Підключення Google Drive

Після виконання цих кроків наш Google Drive буде доступний в папці /content/drive у середовищі Colab, і ми зможете легко звертатися до файлів у Google Drive зі свого коду.

Вкажемо шлях до одного з наших датасетів та отримаємо інформацію про перші декілька рядків датасету:

```
diabetes_df = pd.read_csv('diabetes.csv')

diabetes_df.head()
```

Рис 3.12. Команда отримання даних датасету

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Рис 3.13. Перші 5 даних з датасету 1

Отримаємо інформацію про кожен з атрибутів датасету: номер стовпчика, його назва, кількість рядків та тип даних стовпчику.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Рис 3.14. Інформація про атрибути датасету

Тепер переглянемо детальну інформацію про кожен з атрибутів: кількість значень, мінімальне значення, максимальне значення, та середнє арифметичне.

```
diabetes_df.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Рис 3.15. Детальна інформація про датасет

3.2.3 Підготовка датасетів.

Для початку розділимо датасет на атрибути, що вказають на показники здоров’я, та мітки, що визначають наявність цукрового діабети чи його відсутність.

```
X = diabetes_df.drop('Outcome', axis=1)
y = diabetes_df['Outcome']
```

Рис 3.16. Розділення датасету на атрибути та мітки

Для оптимізації та пришвидшення навчання моделі слід провести шкалювання (Scaling) вхідних ознак. Воно також важливе для того, щоб

нівелювати вплив великих значень, які можуть домінувати над значно меншими значенням та перекривати їх вплив. Основна ціль шкалювання – приведення атрибутів до спільного масштабу. Є декілька видів шкалювання і всі вони наявні у бібліотеці sclearn:

1. Нормалізація (Min-Max Scaling)
   - Опис: приводить значення вхідних ознак до інтервалу від 0 до 1, де 0 – найбільше значення колонки, а 1 – найбільше.
   - Використання: даний тип шкалювання зазвичай використовується тоді, коли для нас важливе абсолютне значення ознак та коли він є нерівномірним.
2. Стандартизація (Z-Scope Scaling)
   - Опис: приводить значення вхідних ознак до інтервалу з середнім значенням 0 та стандартним його відхиленням, що дорівнює 1, тобто кожне значення атрибуту є відношенням різниці між його значенням та середнім значенням колонки до стандартного відхилення ознаки.
   - Використання: даний тип шкалювання використовується, коли розподіл значень ознак є нормальним.

У нашому випадку буде використовуватись стандартизація, оскільки розподіл ознак є нормальним.

```
standard_scaler_X = StandardScaler()
X =  pd.DataFrame(standard_scaler_X.fit_transform(X), columns=['Pregnancies', 'Glucose',
                                                               'BloodPressure', 'SkinThickness',
                                                               'Insulin', 'BMI',
                                                               'DiabetesPedigreeFunction', 'Age'])
X.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.639947	0.848324	0.149641	0.907270	-0.692891	0.204013	0.468492	1.425995
1	-0.844885	-1.123396	-0.160546	0.530902	-0.692891	-0.684422	-0.365061	-0.190672
2	1.233880	1.943724	-0.263941	-1.288212	-0.692891	-1.103255	0.604397	-0.105584
3	-0.844885	-0.998208	-0.160546	0.154533	0.123302	-0.494043	-0.920763	-1.041549
4	-1.141852	0.504055	-1.504687	0.907270	0.765836	1.409746	5.484909	-0.020496

Рис 3.17. Приклад проведення шкалювання та результат його виконання

Тепер нам необхідно розділити дані наявні у датасеті на дані, що будемо використовувати для навчання та дані для тестування. Вони нам потрібні для того, щоб після навчання моделі перевірити її точність.

Для розділення використаємо функцію test_train_split з бібліотеки sklearn.model_selecting:

- Призначення: розділення набору даних на навчальний і тестовий. Це важливо для оцінки продуктивності моделей машинного навчання на невідомих даних.
- Використання: передаються ознаки і цільові змінні, вказується розмір тестового набору (або тренувального набору) і, за необхідності, задається випадкове значення для повторюваності. Функція повертає чотири набори: навчальні ознаки(X_train), тестові ознаки(y_train), цільові значення навчання(X_test) та цільові значення тестування(y_test).

```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.33,
                                                    random_state=7)
```

Рис 3.18. Розділення даних датасету на дані для навчання та тестування

3.2.4 Створення моделей.

Приступимо до створення моделей навчання. Використаємо декілька методів машинного навчання, а саме: Random Forest, Decision Tree, Support Vector Machine. Потім порівняємо їх точність і виберемо ту, яка найкраще підходить до вирішення нашої пробелеми.

Таблиця 3.2.

Опис методів машинного навчання

Назва моделі	Тип	Опис
Random Forest	Ансамблевий	Це ансамбль рішень, який використовує багато рішень (дерев рішень) для прогнозування. Кожне дерево обчислює прогноз, а остаточний прогноз даного методу робиться шляхом підсумовування прогнозів усіх дерев. Важливою особливістю є те, що кожне дерево будується на випадковій

Таблиця 3.2. (Продовження)

Опис методів машинного навчання

		підмножині даних і враховує лише підмножину специфічних особливостей, що допомагає зменшити надмірне припасування і підвищити надійність моделі.
Decision Tree	Класифікація або регресія	Дерева рішень структуровані як дерево, де кожен вузол розділяє вибірку даних за певною ознакою. Кожен лист дерева відповідає кінцевому класу (у випадку класифікації) або числу (у випадку регресії). Вузли розділені відповідно до таких критеріїв, як ентропія або коефіцієнт Джині. Дерева рішень дуже легко інтерпретувати, але вони дуже схильні до однієї розповсюдженої проблеми, яка зустрічається у машинному навчанні, - перенавчання
Support Vector Machine	Класифікація або регресія	Метод опорних векторів створює гіперплощину, яка максимально відокремлює класи. Вектори на межі рішення називаються опорними векторами; Моделі, що створені з використання даного методу, мають велику потужність, коли мають справу з даними високої розмірності, і добре узагальнюють дані. Важливим моментом є вибір відповідної функції ядра, яка визначає, як вимірюються відстані між точками в просторі.

3.2.5 Random Forest.

Спершу створимо модель, яка використовує метод **Random Forest.** Імпортуємо відповідний метод з бібліотеки sklearn.

```
from sklearn.ensemble import RandomForestClassifier

random_forest = RandomForestClassifier(n_estimators=150)
random_forest.fit(X_train, y_train)
```

Рис 3.19. Створення моделі, що використовує метод Random Forest.

При ініціалізації методів нам треба вказати кількість дерев, що буде створено для навчання(n_estimators). Потім передаємо вхідні ознаки(x_train) та мітки.

Після навчання моделі слід перевірити точність моделі як на даних, на яких вона навчалась, так і на даних, які були відсутні. Для цього знадобиться функція accuracy_score, яка перевіряє відповідність прогнозу до істини та вираховує відповідний відсоток точності.

```
from sklearn import metrics

random_forest_train = random_forest.predict(X_train)

print("Accuracy_Score =", format(metrics.accuracy_score(y_train,
                                                        random_forest_train)))
```

```
Accuracy_Score = 1.0
```

Рис 3.20. Перевірка точності методу Random Forest на даних, яких він навчався

```
predictions = random_forest.predict(X_test)

print("Accuracy_Score =", format(metrics.accuracy_score(y_test, predictions)))
```

```
Accuracy_Score = 0.7637795275590551
```

Рис 3.21. Перевірка точності методу Random Forest на нових даних

Можемо бачити, що дана модель має досить високі показники точності як на множині даних, на яких відбувалося тренування, так і на даних, що були завчасно відведені для тестування, а саме 100% та 76% точності відповідно.

3.2.6 Decision Tree

Тепер створимо модель, яка використовує метод Decision Tree. Імпортуємо відповідний метод з бібліотеки sklearn.

```
from sklearn.tree import DecisionTreeClassifier

decision_tree = DecisionTreeClassifier()
decision_tree.fit(X_train, y_train)
```

Рис 3.22. Створення моделі, що використовує метод Random Forest

Перевіримо точність даної моделі використовуючи такий самий алгоритм як і в минулому підрозділі.

```
decision_tree_train = decision_tree.predict(X_train)

print("Accuracy_Score =", format(metrics.accuracy_score(y_train,
                                                        decision_tree_train)))
```

```
Accuracy_Score = 1.0
```

Рис 3.23. Перевірка точності методу Decision Tree на даних, яких він навчався

```
predictions = decision_tree.predict(X_test)

print("Accuracy Score =", format(metrics.accuracy_score(y_test,predictions)))
```

```
Accuracy Score = 0.6968503937007874
```

Рис 3.24. Перевірка точності методу Decision Tree на нових даних

3.2.7 Support Vector Machine.

На останок створимо модель, яка використовує метод Support Vector Machine. Імпортуємо відповідний метод з бібліотеки sklearn

```
from sklearn.svm import SVC

support_vector_machine = SVC()
support_vector_machine.fit(X_train, y_train)
```

Рис 3.25. Створення моделі, що використовує метод Support Vector Machine

Перевіримо точність даної моделі використовуючи такий самий алгоритм як і в минулому підрозділі.

```
prediction = support_vector_machine.predict(X_test)

print("Accuracy Score =", format(metrics.accuracy_score(y_test, prediction)))
```

```
Accuracy Score = 0.7519685039370079
```

Рис 3.26. Перевірка точності методу Support Vector Machine на даних, яких він навчався

Легко бачити, що точність даного методу під час перевірки на даних, який він навчання, суттєво знизилась в порівнянні з попередніми методами.

```
prediction = support_vector_machine.predict(X_test)

print("Accuracy Score =", format(metrics.accuracy_score(y_test, prediction)))
```

```
Accuracy Score = 0.7519685039370079
```

Рис 3.27. Перевірка точності методу Support Vector Machine на нових даних

Висновок до розділу 3

У цьому розділі було проведено опис атрибутів датасетів, огляд бібліотек та розроблено декілька моделей, що використовують різні методи машинного навчання.

Дійшли висновку, що кожен з датасетів має певну множину атрибутів, кожен з яких відповідає за деякий показник здоров’я пацієнту. За допомогою, цих атрибутів можна визначити наявність чи відсутність хвороби цукрового діабету з певною точністю

У процесі розробки було спершу проведено нормалізацію даних з метою максимального пришвидшення навчання моделей. Потім створено декілька моделей, які навчаються на даних датасету та роблять прогнози. Після навчання моделей було оцінено точність кожної з них та проведено порівняльний аналіз. Дійшли висновку, що найбільш точною є модель, яка використовує метод Random Forest.

РОЗДІЛ 4
РОЗРОБКА СТАРТАП ПРОЄКТУ

4.1 Опис ідеї проєкту.

У сучасних реаліях проблеми здоров'я та медицини стають все більш актуальними і вимагають інноваційних підходів. Технології, які доступні нам сьогодні, дозволяють нам робити революцію в сфері діагностування різних розповсюджених захворювань. До таких можна віднести цукровий діабет, який є надзвичайно важливою проблемою і набуває характеру епідемії світового масштабу. Швидка і точна діагностика цього захворювання є запорукою надання ефективного лікування та підтримки задовільного стану людей, що переживають цю недугу.

Метою даного проєкту є розробка автоматизованої системи для діагностування цукрового діабету на основі машинного навчання. Система повинна забезпечити швидку, достатньо точну та доступну діагностику для пацієнтів та медичних працівників. Дана система може використовуватись як допоміжний інструмент в процесі діагностування хвороби цукрового діабету, який дозволить попередньо визначити його наявність з певною ймовірністю опираючись на попередньо відомі показники здоров’я пацієнту. Як наслідок, лікар може надати певний курс лікування, дієту або деякі вказівки, яких хворому бажано було б дотримуватись до кінця повного процесу діагностування хвороби, коли вже буде достовірно ймовірно про наявність чи відсутність цукрового діабету. Тоді вже лікар зможе надати максимально якісну допомогу.

Таблиця 4.1.

Опис ідеї стартап-проєкту

Зміст ідеї	Напрямки застосування	Користь для користувача система
Автоматизована система діагностування цукрового діабету	Машинне навчання	Можливість попередньо визначити наявність захворювання

Таблиця 4.2.

Карта сегментування ринку

	Вил		
	Заміна лікаря	Аналіз роботи лікаря	Отримання попереднього діагнозу
Великі клініки	Державні	Державні	
Середні клініки	Приватні	Приватні	
Малі клініки			Пацієнт

4.2 Технологічний аудит ідеї проєкту.

Тепер слід провести технологічний аудит, тобто короткий огляд технологій, що використовуються для написання.

У наступній таблиці наведено показники за якими проведено аудит та визначено за якими технологіями буде створено продукт.

Таблиця 4.3.

Технологічний аудит

№	Ідея проєкту	Технології і реалізації	Наявність технологій	Доступність технологій
1	Розробка нейронної мережі	Python	+	+
2	Бібляотека для побудови нейронних мереж	sklearn	+	+
3	Сервіс для розробки та хмарних обчислень	Google Colab	+	+

4.3 Аналіз ринкових можливостей для запуску стартап-проекту.

Визначимо ринкові можливості, які можна використати під час ринкового впровадження проекту та ринкових загроз, які можуть зашкодити реалізації проекту.

Таблиця 4.4.

Аналіз попиту та стан ринку

№	Показники ринку	Характеристика
1	Кількість гравців	Відсутні
2	Динаміка ринку	Зростає
3	Наявність обмежень для вхожу	Відсутні
4	Вимоги до стандартизації та сертифікації	Відсутні

4.4 Аналіз потенційних клієнтів.

Тепер слід провести аналіз потенційних клієнтів, які можуть використовувати дану систему

Таблиця 4.5.

Характеристика потенційних клієнтів стартап-проєкту

№	Цільова аудиторія	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Великі клініки	Допомога у діагностуванні хвороби	Точність та низька вартість
2	Маленькі клініки	Бажання підтримки конкурентспроможності	Точність та простота використання
3	Фізичні особи	Можливість самостійно оцінинити наявність хвороби	Точність

Тепер слід визначити фактори загроз, з якими може зіткнутися компанія у майбутньому.

Таблиця 4.6.

Фактори загроз

№	Фактор	Зміст загрози	Можлива реакція
1	Поява більш швидких та точних алгоритмів	Поява нового більш швидкого та точного алгоритму може зменшити попит на наш алгоритм.	Зміна моделі штучного інтелекту на більш досконалу, що матиме більшу точність та швидкість.
2	Відсутність інвестицій	Зменшення продуктивності розробників, через що у майбутньому може бути програна конкуренція на ринку	Пошук додаткових інвестицій

Таблиця 4.6. (Продовження)

Фактори загроз

3	Збільшення конкуренції	Оскільки штучний інтелект є технологією, яка має великий попит сьогодні, постійно буду з'являтися нові гравці, з якими необхідно буде конкурувати.	Пошук нових більш точних рішень.

Таблиця 4.7.

Фактори можливостей

№	Фактор	Зміст можливості	Можлива реакція
1	Інвестиції	Отримання додаткових коштів для фінансування проєкту	Вкладання інвестицій в розробку та залучення нових розробників
2	Отримання контракту	Підписання контракту з великою компанією	Збільшення обсягів виробництва
3	Покращення існуючого методу	Збільшення конкурентоспроможності, що дозволить зробити даний метод більш привабливим.	Збільшення популярності

Одним з важливих етапів є проведення аналізу пропозицій та визначення рівню конкуренції.

Таблиця 4.8.

Ступеневий аналіз конкуренції на ринку

№	Особливості конкурентного середовища	Зміст даної характеристики	Вплив на діяльність проєкту
1	Чиста конкуренція	Наявність у відкритому доступі великої кількості алгоритмів та бібліотек	Створення власних алгоритмів, які не матимуть альтернативи
2	Міжнародна конкуренція	Відсутність прив'язки до певної географічної місцевості	Можливість використання проєкту у різних країнах та збільшення його популярності
3	Товарно-родова конкуренція	Наявність багатьох товарів однієї категорії	Необхідність проведення більш активної рекламної компанії ніж конкуренти
4	Конкурентні переваги компанії	Клієнти готові витрачати більші кошти за більш якісний продукт	Використання отриманих доходів на розвиток компанії
5	Марочна конкуренція	Система призначена для спільної цільової аудиторії	Зосередження уваги клієнта на бренді.

Також проведемо аналіз конкуренції в галузі за моделлю 5 сил М. Портера. Даний метод є досить популярним, та допомагає якісно визначити доречність певного товару для виходу на ринок.

Таблиця 4.9.

Аналіз конкуренції в галузі за моделлю М. Портера

Складові аналізу	Прямі конкуренти	Потенційні конкуренти	Постачальники	Клієнти	Замінники
	Товари, що використовують альтернативні методі	Немає	Визначити фактори сили постачальників	Слід визначити фактори, які найбільш необхідні для споживача	Немає
Висновки	Конкурента боротьба, де необхідно довести перевагу машинного навчання	Можливість виходу на ринок різних конкурентів	Відсутні	Слід досягати максимальної якості роботи проєкту	У замінниках може бути використання більш дешева технологія

Визначимо фактори конкурентоспроможності:

Таблиця 4.10.

Обґрунтування факторів конкурентоспроможності

№	Фактор конкурентоспроможності	Обґрунтування
1	Низька собівартість	Можливо повторне використання моделі для вирішення схожих задач
2	Інноваційність	Створення нових технологій, які раніше не використовувалися, завжди приваблюють інвесторів

Після визначення факторів конкурентоспроможності проведемо аналіз сильних та слабких сторін стартап проєкту.

Таблиця 4.11.

Порівняльний аналіз сильних та слабких сторін проєкту

№	Фактор конкурентоспроможності	Бали 1-20	Рейтинг конкурентів у порівнянні з власним проєктом						
			-3	-2	-1	0	+1	+2	+3
1	Міжгалузевість					+			
2	Інноваційність						+		
3	Низька собівартість				+				
4	Простота експлуатацій								+

4.5 SWOT-аналіз стартап-проєкту.

Фінальним етапом ринкового аналізу можливостей є проведення SWOT-аналізу, Цей інструмент допомагає проаналізувати внутрішні фактори (сильні(1 чверть) та слабкі(2 чверть) сторони), які впливають, і зовнішні фактори (можливості(3 чверть) та загрози(4 чверть)), які можуть мати вплив на організацію.

Таблиця 4.12.

SWOT-аналіз стартап проєкту

	Корисні	Шкідливі
Внутрішні	**Сильні сторони**	**Слабкі сторони**
	Низька ціна	Наявність багатьох схожих систем, що наразі у розробці
	Висока швидкість	Необхідність постійної модернізації
	Досить висока точність	
	Можливість покращення системи за рахунок оновлення моделі	

Таблиця 4.12. (Продовження)

SWOT-аналіз стартап проєкту

Зовнішні	**Можливості**	**Погрози:**
	Побудова системи, що навчається на різних даних	Вихід на ринок компанії, які мають більший ресурс для розробки кращих систем
	Залучання інвесторів	Використання тільки для нестандартних задач

Тепер проведемо оцінку термінів та ймовірність отримання ресурсів.

Таблиця 4.13.

Терміни та ймовірність отримання ресурсів

№	Орієнтований комплекс заходів	Ймовірність отримання ресурсів	Строки реалізацій
1	Розробка продукту, просування бренду	70%	10 місяців
2	Розробка продукту та його безкоштовне розповсюдження	35%	6 місяців

Висновок до розділу 4

В результаті виконання даного розділу, де було проведено аналіз ринку, його стан та можливий попит на даний продукт, можна дійти висновку, що він має потенціал для успішної реалізації на ринку, через велику кількість потенційних клієнтів та попит на технології, що використовують алгоритми машинного навчання.

Важливо відзначити можливу велику прибутковість даного продукту, навіть враховуючи високу вартість на отримання ліцензії для виходу на ринок. Це можливо завдяки низькій конкуренції на ринку сьогодні та його конкурентоспроможність.

ВИСНОВОК

Як і було зазначено при виконанні даної роботи, у сьогоднішньому світі є безліч хвороб, які швидко поширюються серед населення нашої планети. Причиною цьому є як і сучасний спосіб життя людей, їх харчування, так і такі зовнішні чинники як якість екології, відсутність якісної медицини у нерозвинених країнах, тощо.

На допомогу приходить машинне навчання, яке поступово стає невід'ємною частиною нашого життя та стає використовуватись у абсолютно різних сферах: від звичайного голосового помічника до алгоритмів, які мають змогу повністю замінити людину, наприклад деякі сучасні автомобілі можуть без проблем здійснювати пересування дорогами загального користування без втручання водія. Це говорить про неймовірні можливості машинного навчання, які тяжко навіть сьогодні реально оцінити.

Сфера медицини не стоїть осторонь та також починає працювати у напрямку використання неронних алгоритмів. Вони мають змогу значно полегшити роботу лікарів, а й іноді, в деяких випадках, повністю замінити їх.

У даній роботі було проведено аналіз хвороби цукрового діабету та перевірено доречність використання штучного інтелекту для її діагностування. Можна дійти висновку, що дані алгоритми мають неймовірний потенціал у цій сфері, отримуючи досить гарні результати точності, використовуючи навіть доволі прості нейронні алгоритми. Для отримання максимально високої точності та початку використання алгоритмів машинного навчання у медицині необхідно ще провести тяжку роботу у створенні правильних датасетів та нейронних алгоритмів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Berestizhevsky S. Supervised Machine Learning. Chapman and Hall/CRC, 2020.
2. Celebi M. E., Aydin K. Unsupervised Learning Algorithms. Springer, 2018. 558 p.
3. MARSH T. Diabetes: Types, Treatment and Dietary Plan. Independently Published, 2022.
4. Mohan V., Rao G. H. Type 2 Diabetes in South Asians: Epidemiology, Risk Factors & Prevention. Jaypee Brothers Medical Publishers (P) Ltd., 2007.
5. Purdy C. W. 1.-1. Diabetes: Its Causes, Symptoms, and Treatment. Creative Media Partners, LLC, 2018.
6. Reinforcement Learning / ed. by M. Wiering, M. van Otterlo. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012.
7. Shaw K. Diabetes: Chronic complications. 3rd ed. Chichester, West Sussex, UK : John Wiley & Sons, 2012.
8. Slavio J. Neural Networks: Neural Networks Tools and Techniques for Beginners. John Slavio, 2019. 114 p.
9. Textbook of Diabetes / B. J. Goldstein et al. Wiley & Sons, Incorporated, John, 2016. 1104 p.
10. Ventriglia F. Neural Modeling and Neural Networks. Elsevier Science & Technology Books, 2013.
11. Python Language Advantages and applications. [Електронний ресурс] / GeeksForGeeks – Режим доступу до ресурсу: https://www.geeksforgeeks.org/python-language-advantages-applications/
12. TensorFlow. [Електронний ресурс] / TensorFlow – Режим доступу до ресурсу: https://www.tensorflow.org/
13. Keras. [Електронний ресурс] / Keras – Режим доступу до ресурсу: https://keras.io

14. NumPy. [Електронний ресурс] / NumPy – Режим доступу до ресурсу: https://numpy.org/.
15. 4 types of learning in machine learning explained [Електронний ресурс] / TechTarget – Режим доступу до ресурсу: https://www.techtarget.com/searchenterpriseai/tip/Types-of-learning-in-machine-learning-explained
16. What is blood preasure? [Електронний ресурс] / NHS – Режим доступу до ресурсу: https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/#:~:text=systolic%20pressure%20–%20the%20pressure%20when,your%20heart%20rests%20between%20beats
17. Insulin. [Електронний ресурс] / Cleveland Clinic – Режим доступу до ресурсу: https://my.clevelandclinic.org/health/articles/22601-insulin
18. Body mass index is a more powerful risk for diabetes than genetics. [Електронний ресурс] / BHF – Режим доступу до ресурсу: https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2020/august/body-mass-index-is-a-more-powerful-risk-factor-for-diabetes-than-genetics
19. Higher Numbers of Pregnancies Associated With an Increased Prevalence of Gestational Diabetes Mellitus: Results From the Healthy Baby Cohort Study. [Електронний ресурс] / NLM – Режим доступу до ресурсу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153959/
20. Machine learning in precision diabetes care and cardiovascular risk prediction. [Електронний ресурс] / BMC – Режим доступу до ресурсу: https://cardiab.biomedcentral.com/articles/10.1186/s12933-023-01985-3