Міністерство освіти і науки України
Національний технічний університет
«Дніпровська політехніка»

Факультет інформаційних технологій
(факультет)

Кафедра системного аналізу та управління
(повна назва)

ПОЯСНЮВАЛЬНА ЗАПИСКА
кваліфікаційної роботи ступеня магістра

Студентки Сидоренко Катерини Віталіївна

академічної групи 124м-22-1

спеціальності 124 Системний аналіз

на тему: «Аналіз причин та прогнозування виявлення цукрового діабету методами машинного навчання»

Керівники	**Прізвище, ініціали**	**Оцінка за шкалою**		**Підпис**
		рейтинговою	**інституційною**	
кваліфікаційної роботи	*к.ф.-м.н., доц. Хом'як Т.В.*			
розділів:				
Інформаційно-аналітичний	*к.ф.-м.н., доц. Хом'як Т.В.*			
Спеціальний	*к.ф.-м.н., доц. Хом'як Т.В.*			
Рецензент				
Нормоконтролер	*к.ф.-м.н., доц. Хом'як Т.В.*			

Дніпро
2023

ЗАТВЕРДЖЕНО:

завідувач кафедри

Системного аналізу та управління
(повна назва)

_____ *к.т.н., доц. Желдак Т.А.*
(підпис) (прізвище, ініціали)

«_____»_____20___року

ЗАВДАННЯ
на кваліфікаційну роботу
ступеня *магістра*

студентці *Сидоренко К.В.* **академічної групи** *124м-22-1*
спеціальності: *124 Системний аналіз*
на тему *«Аналіз причин та прогнозування виявлення цукрового діабету методами машинного навчання»*
затверджену наказом ректора НТУ «Дніпровська політехніка»
від *18.05.2022 р. №268-с*

Розділ	Зміст	Термін виконання
1. Інформаційно-аналітичний розділ	*Опис технологій штучного інтелекту, зокрема машинного навчання, його методи та складові, оцінка якості моделей.*	01.07.2023 – 31.08.2023
2.Спеціальний розділ	*Аналіз причин виявлення цукрового діабету у пацієнтів, прогнозування методами машинного навчання: Decision Tree, Random Forest, K-NN, Ada Boost та вибору найефективнішого з них.*	01.09.2023 – 01.11..2023

Завдання видано _____ *доц. Хом'як Т.В.*
(підпис) (прізвище, ініціали)

Дата видачі: *01.10.2022 р.*

Дата подання до екзаменаційної комісії: _____

Прийнято до виконання _____ Сидоренко К. В.
(підпис студента) (прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 81 с., 30 рисунків,6 таблиць, 34 джерела.

Об'єкт дослідження: є процес реалізації програмного продукту для розв'язання задачі прогнозування й аналізу причин виникнення цукрового діабету до його появи аби запобігти його розвитку.

Предмет дослідження: методи машинного навчання Decision Tree, Random Forest, Ada Boost, K-NN.

Метою даної кваліфікаційної роботи є аналіз причин за допомогою візуалізації та прогнозування виникнення діабету методами машинного навчання для того, щоб зменшити кількість випадків виникнення діабету.

Методи дослідження: методи попереднього опрацювання даних: Isolated Forest (видалення викидів), SMOTEENN (балансування даних), Standard Scaler (масштабування) та методи машинного навчання – Decision Tree, Random Forest, Ada Boost, K-NN.

В інформаційно-аналітичному розділі було проаналізовано результати досліджень науковців, які були здійснені раніше, обрано методи машинного навчання за допомогою яких буде здійснено прогнозування та критерії оцінки якості моделі класифікації.

У спеціальному розділі було проведено аналіз початкових даних, а саме: видалення викидів, розподілення даних, масштабування та балансування даних, проаналізовано вплив факторів на виявлення цукрового діабету до його появи й зроблено прогнозування чотирма методами машинного навчання для виявлення діабету та обрано найефективніший з них.

Практична цінність отриманих у роботі результатів полягає у створенні більш точної моделі для прогнозування виявлення цукрового діабету, виявленні значно більшої кількості випадків діабету до його появи, ефективнішому та ранньому лікуванню, зменшенні витрат на медичне обслуговування.

Ключові слова: МАШИННЕ НАВЧАННЯ, ЦУКРОВИЙ ДІАБЕТ, ПРОГНОЗУВАННЯ, ОБРОБКА ДАНИХ, DECISION TREE, Random Forest, Ada Boost, K-NN, МОДЕЛЬ.

ABSTRACT

Explanatory note: 81 p., 30 figures, 6 tables, 34 sources.

The object of the research: there is a process of implementing a software product to solve the problem of forecasting and analyzing the causes of diabetes before its appearance in order to prevent its development.

Research subject: methods of machine learning Decision Tree, Random Forest, Ada Boost, K-NN.

The purpose of this qualification is to analyze the causes using visualization and predict the occurrence of diabetes using machine learning methods in order to reduce the number of diabetes occurrences.

Research methods: data preprocessing methods: Isolated Forest (outlier removal), SMOTEENN (data balancing), Standard Scaler (scaling) and machine learning methods - Decision Tree, Random Forest, Ada Boost, K-NN.

In the informational and analytical section, the results of research by scientists that were carried out earlier were analyzed, machine learning methods were selected, which will be used to make predictions and criteria for assessing the quality of the classification model.

In a special section, the analysis of the original data, namely: removal of outliers, data partitioning, scaling and balancing of data, analyzed the influence of factors on the detection of diabetes before its onset, and made predictions by four machine learning methods for diabetes detection and selected the most effective one.

The practical value of the results obtained in the work consists in creating a more accurate model for predicting the detection of diabetes, detecting a significantly greater number of cases of diabetes before its appearance, more effective and early treatment, and reducing the cost of medical care.

Keywords: MACHINE LEARNING, DIABETES DIABETES, PREDICTION, DATA PROCESSING, DECISION TREE, Random Forest, Ada Boost, K-NN, MODEL.

ЗМІСТ

ВСТУП

Сьогодні кількість людей, які живуть з невиліковними хворобами зростає. Цукровий діабет - це серйозне захворювання, яке може призвести до численних ускладнень та проблем зі здоров'ям. Іноді люди через ускладнення діабету, які не контролюється помирають, а саме може статися інфаркт, гіпоглікемія та інші.

Зараз цукровий діабет є однією з найпоширеніших хронічних захворювань у світі, яким страждає близько 530 мільйонів людей, з яких 1 300 000 – громадяни України на червень 2023 року. Це захворювання впливає на рівень цукру (глюкози) в крові. Поява цукрового діабету зазвичай обумовлена генетичними, середовищними та стильовими факторами. Основні причини включають генетичну схильність, ожиріння, неправильну харчову поведінку, інсулінорезистентність та шкідливі звички.

Значущим є той факт, що вчасне виявлення захворювання може запобігти його розвитку. Багато симптомів цукрового діабету, таких як сухість у роті, часті сечовипускання, погіршення зору, втрата ваги, постійне відчуття голоду, не завжди відразу розглядаються як ознаки захворювання. Важливо підкреслити, що ці симптоми можуть бути ранніми показниками високого рівня глюкози у крові.

Отже, для значно більшої ймовірності виявлення цукрового діабету до його появи потрібно мати не тільки більш досвідчених лікарів, а й навчитися прогнозувати дану хворобу для того, щоб у майбутньому не збільшувалась кількість захворюваннях у рік.

Задача даної кваліфікаційної роботи полягає в аналізі факторів, які впливають на ризик розвитку цукрового діабету та прогнозуванні методами машинного навчання, такі як Decision Tree, Random Forest, k-NN та Ada Boost.

Економічний ефект від отриманих результатів дасть змогу виявляти значно більшу кількість випадків діабету до його появи, ефективніше та раннє лікування, зменшити витрат на медичне обслуговування.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ШІ – штучний інтелект

МН – машинне навчання

RF – Random Forest

K-NN – K-Nearest Neighbour

ІХС – ішемічна хвороба серця

АТ – артеріальний тиск

ІМТ – індекс маси тіла

1. ІНФОРМАЦІЙНО-АНАЛІТИЧНИЙ РОЗДІЛ

1.1 Штучний інтелект

Сьогодні штучний інтелект є одним з найпопулярніших термінів у світі. За деякими прогнозами до 2035 року ШІ принесе світовій економіці 15,7 трильйона доларів. Він вже створює купу цифрового контенту — тексти, картинки, музику, відео тощо. Основи його функціонування варто знати хоча б для того, щоб не втратити бізнес чи роботу.

Штучний інтелект (ШІ) — це здатність машин симулювати розум та імітувати людські когнітивні здібності. Тобто збирати й адаптувати зовнішні дані, а на їх основі навчатися ухвалювати рішення та робити висновки, як могла би людина.

Ця галузь включає в себе різноманітні методи та підходи, включаючи машинне навчання, глибоке навчання, обробку природних мов, комп'ютерний бачення, робототехніку та багато інших. Штучний інтелект використовується в різних сферах, включаючи медицину, автономні автомобілі, фінанси, ігри, аналітику даних, мовленнєві асистенти і багато інших областей.

Основна мета ШІ - створити комп'ютерні системи, які можуть розв'язувати складні завдання і адаптуватися до змін в навколишньому середовищі, щоб покращити якість життя та ефективність різних сфер людської діяльності.

Поняття штучного інтелекту нерозривно пов'язане з рядом термінів:

- Big data: щоб навчатися та розвиватися, ШІ потребує колосальних обсягів

даних.

- Machine learning, або машинне навчання, — підгалузь ШІ, яка тренує машини самостійно виконувати завдання на основі отриманих даних, а не просто діяти згідно з інструкціями.
- Deep learning, або глибинне навчання, — підрозділ машинного навчання,

який за допомогою нейромереж тренує машину сприймати великі обсяги необроблених даних (текст, аудіо, відео тощо) та зважати на них.

- Natural language processing, або опрацювання природної мови, — напрям ШІ, який аналізує текстові, аудіо- та відеодані, щоб навчитися синтезувати та імітувати живе людське спілкування.
- Нейромережа — обчислювальна система, яка навчається шляхом аналізу прикладів і поступово покращує свої здібності. Нейромережа діє за принципом центральної нервової системи. [2]

У наш час потенціал застосування штучного інтелекту дуже широкий. Зараз ШІ використовується у багатьох сферах, а саме:

- Медицина;
- Фінанси;
- Промисловість;
- Торгівля;
- Робототехніка;
- Побут людини.

Також неабияку роль відіграє штучний інтелект у роботі підприємств. Він допомагає автоматизувати процеси, які потребують неабияких зусиль, і тоді участь людини залишається мінімальною. [3]

Перспективи штучного інтелекту неабиякі: підвищення ефективності, зручність, позбавлення довготривалих процесів і автоматизація звичних. Поки що порівняно новий напрям стикається з низкою труднощів щодо впровадження рішень в життя. [4]

Штучний інтелект (ШІ) має численні переваги та недоліки.

Переваги штучного інтелекту:

- точність в обробці даних;
- здатність аналізувати велику кількість інформації з великою швидкістю;
- ШІ не потрібен сон і перерва на обід, він не допускає помилок через перевтому;

- може вирішувати завдання, які для людини були б занадто складними

або неможливими через обробку великих обсягів даних.

- може бути використаний для автоматизації багатьох рутинних

завдань у бізнесі та промисловості, що призводить до збільшення продуктивності.

- використовувати штучний інтелект можна там, де людині небезпечно

перебувати.

Недоліки штучного інтелекту:

- обмеження в загальному розумінні;
- потреба в великих обсягах даних для якісного навчання;
- вразливість до атак;
- втрата людських робочих місць.

Отже, ШІ має значні переваги у багатьох сферах, але важливо враховувати його обмеження та питання, пов'язані з етикою та безпекою, при використанні цієї технології.

1.2 Машинне навчання

Машинне навчання (МН) або Machine learning – один з розділів AI, алгоритми, що дозволяють комп'ютеру робити висновки на підставі даних, не слідуючи жорстко заданим правилами. Тобто машина може знайти закономірність у складних і багато-параметричних завданнях (які мозок людини не здатен вирішити), таким чином знаходячи більш точні відповіді. Як результат – правильне прогнозування. [5]

Ціль машинного навчання – частково або й повністю автоматизувати рішення різних складних аналітичних задач.

Тому, насамперед, машинне навчання покликане давати максимально точні прогнози на підставі вступних даних, щоб власники бізнесів, маркетологи

і співробітники могли приймати правильні рішення в своїй роботі. В результаті навчання машина може передбачати результат, запам'ятовувати його, відтворювати за необхідності, вибирати кращий із декількох варіантів.

На даний момент машинне навчання охоплює широкий спектр додатків від банків, ресторанів, заправок до роботів на виробництві. Нові завдання, що виникають практично щодня, призводять до появи нових напрямків машинного навчання.

Складові машинного навчання

Наразі, для якісного машинного навчання потрібно три речі: дані, ознаки та алгоритми. Схема складових машинного навчання зображена на рисунку 1.1.

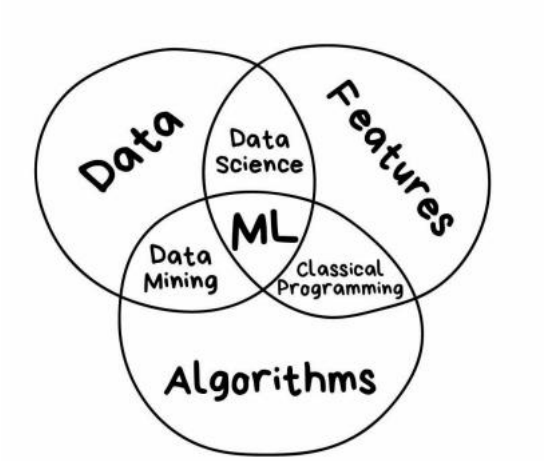


Рисунок 1.1 – Складові машинного навчання

Дані є основним ресурсом у машинному навчанні. Вони можуть бути представлені у різних форматах, таких як текстові дані, зображення, аудіо, відео, числові дані і т. д. Для успішного навчання моделі необхідно мати якісні та репрезентативні дані. Важливо збирати, очищувати та підготовлювати дані перед їх використанням у моделі. Головні критерії:

- важливість даних: дані є основними будівельними блоками машинного навчання. Якість та кількість даних впливають на якість моделі. Недостатні або забруднені дані можуть призвести до поганої роботи моделі;
- збір та збереження даних: для початку роботи з даними їх необхідно зібрати і зберегти відповідним чином. Це може включати в себе роботу з базами даних, веб-скрапінг, збір даних вручну або через датчики;

- очищення та підготовка даних: дані часто потребують очищення від викидів (outliers), відсутніх значень (missing values) та інших аномалій. Також їх може бути необхідно перетворити або закодувати для подальшого використання в моделі;

- розподіл даних: Дані зазвичай поділяють на навчальний (training), валідаційний (validation) та тестовий (testing) набори для навчання, налаштування та оцінки моделі відповідно.

Ознаки (фічі) представляють собою характеристики або атрибути даних, які модель використовує для прийняття рішень. Вони можуть бути числовими (наприклад, розмір, вага, температура) або категоріальними (наприклад, кольори, класи, типи). Вони можуть бути одномірними, багатовимірними або навіть текстовими. Головні критерії:

- види ознак: ознаки можуть бути числовими, категоріальними, текстовими, географічними тощо. Різні типи ознак вимагають різних методів обробки та інженерії фіч;

- важливість ознак: деякі ознаки можуть мати більший вплив на вихід моделі, ніж інші. Аналіз важливості ознак допомагає вибрати найбільш значущі фактори для моделі;

- інженерія фіч – це процес створення нових ознак на основі існуючих даних, що може покращити роботу моделі. Наприклад, створення поліноміальних ознак або витягування додаткової інформації з тексту. [7]

Алгоритми в машинному навчанні визначають, як модель аналізує дані та вчиться з них. Існує багато різних алгоритмів машинного навчання, які можуть бути використані в залежності від завдання. До них входять:

- навчання з учителем (supervised learning): Навчання класифікації (classification) та регресії (regression);

- навчання без учителя (unsupervised learning): Кластеризація (clustering) та зменшення розмірності (dimensionality reduction);

- навчання з підкріпленням (reinforcement learning): Навчання агентів приймати дії на основі нагород та покарань. [5]

У процесі машинного навчання, алгоритми використовують дані та їх ознаки для навчання моделей, які можуть робити прогнози або приймати рішення на основі нових даних. Після навчання модель може бути використана для розв'язання різних завдань, таких як класифікація, прогнозування, аналіз або управління. Успішне використання машинного навчання вимагає ретельної роботи з даними, вибору відповідних ознак та алгоритмів, а також налаштування та оцінки моделі.

Основні етапи машинного навчання включають в себе:

- перший крок у машинному навчанні – збір відповідних даних, які будуть використовуватися для навчання алгоритму машинного навчання. Це можна зробити за допомогою скрапінгу веб-сайтів, збору даних з датчиків пристроїв Інтернету речей або ручного введення даних в базу даних;
- після збору дані потрібно очистити та підготувати для аналізу. Цей процес включає видалення будь-яких відсутніх або непотрібних даних, перетворення даних до формату, який може використовувати алгоритм, та розбиття даних на набори (сети) для навчання та тестування;
- на цьому етапі алгоритм машинного навчання тренується на підготовлених

даних. Цей процес включає вибір відповідного алгоритму та подачу йому навчальних даних, що дозволяє алгоритму вивчити закономірності в даних та налаштувати свої параметри для мінімізації помилок;

- після того, як модель навчена, її потрібно випробувати на тестових даних для оцінки її продуктивності. Процес включає прогнозування результатів на тестових даних та порівняння їх з фактичними результатами.
- після успішного тестування модель можна впровадити для прогнозування нових даних. Це можна зробити шляхом інтеграції моделі в існуючу програму або створення нової програми (застосунку), яка використовує цю модель. [6]

1.3 Методи машинного навчання

Зараз існує багато методів машинного навчання, а саме основні методи:

- класичне навчання - відомі і добре вивчені алгоритми навчання, розроблені в основному більше 50-ти років тому для статистичних бюро. Підходить, насамперед, під завдання роботи з даними: класифікація, кластеризація, регресія і т.п. Застосовують для прогнозування, сегментації клієнтів і так далі;
- нейронні мережі і глибоке навчання - найбільш сучасний підхід до машинного навчання. Нейронні мережі застосовуються там, де потрібні розпізнавання або генерація зображень і відео, складні алгоритми управління або прийняття рішень, машинний переклад і подібні складні завдання;
- навчання з підкріпленням використовують там, де перед машиною стоїть завдання - правильно виконати поставлені завдання у зовнішньому середовищі маючи безліч можливих варіантів дії. Наприклад в комп'ютерних іграх, трейдингових операціях, для безпілотної техніки;
- методи ансамблю (Ensemble Methods)- це комбінація кількох моделей для отримання кращої загальної продуктивності. Приклади включають в себе випадковий ліс (Random Forest), бустінг (Boosting) і багатокласову класифікацію (Multi-class classification). [5]

Класичне навчання включає у себе наступні навчання:

- навчання з учителем застосовують коли потрібно навчити машину розпізнавати об’єкти або сигнали. Загальний принцип навчання з учителем це «дивись, ось це двері і це теж двері, і ось це теж двері»;
- навчання без вчителя використовує принцип «ця річ така ж як інші». Алгоритми вивчають подібності і можуть виявити відмінність і виконати виявлення аномалій, розпізнаючи, що є незвичайним або несхожим.

Класифікація методів машинного навчання зображена на рисунку 1.2.

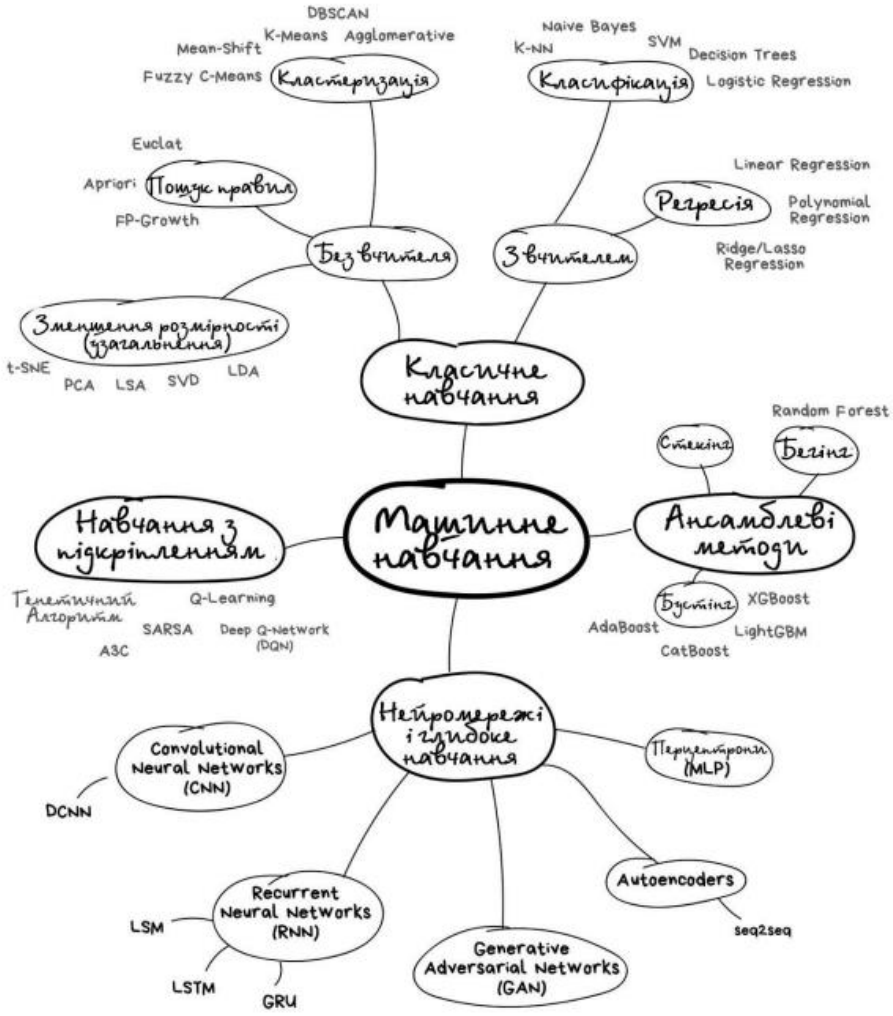


Рисунок 2.2 – Класифікація методів машинного навчання

Навчання з учителем (Supervised Learning):

- класифікація (Classification): Модель навчається визначати класи або категорії для нових даних. Приклади включають логістичну регресію, метод опорних векторів (SVM), нейронні мережі та багато інших.
- регресія (Regression): Модель навчається прогнозувати числові значення на

основі даних. Приклади включають лінійну регресію, дерева рішень та ансамблі рішень.

У даній кваліфікаційній роботи буду використане класичне навчання з вчителем за допомогою класифікації. Тому далі буде більш детально описано про методи класифікації машинного навчання, а саме про методи: дерево рішень, випадковий ліс, k-найближчих сусідів, AdaBoost.

1.3.1 Оцінка якості моделі класифікації

Матриця помилок (Confusion Matrix) – це таблиця, яка дозволяє візуалізувати результати класифікації моделі в машинному навчанні. Вона допомагає розраховувати кількість правильних і неправильних класифікацій для кожного класу, допомагаючи зрозуміти, наскільки ефективно модель вирішує завдання класифікації. Матриця помилок використовується для обчислення різних метрик якості, таких як точність, повнота, F-мера і інші.

Матриця помилок складається з чотирьох основних елементів:

- True Positives (TP): це кількість прикладів, які були правильно класифіковані як позитивні. Іншими словами, це кількість правильних передбачень, що належать до позитивного класу.
- False Positives (FP): це кількість прикладів, які були неправильно класифіковані як позитивні. Це означає, що модель вказала на позитивний клас, коли він фактично належав до негативного класу.
- True Negatives (TN): це кількість прикладів, які були правильно класифіковані як негативні. Тобто це кількість правильних передбачень для негативного класу.
- False Negatives (FN): це кількість прикладів, які були неправильно класифіковані як негативні. Це означає, що модель вказала на негативний клас, коли він фактично належав до позитивного класу. [8]

Матриця помилок, одна з найважливіших речей, на яку потрібно дивитися при оцінці моделі класифікації. Це матриця, яка візуалізує кількість фактичних екземплярів класу в порівнянні з прогнозованими екземплярами класу. Таке

подання дозволяє нам швидко побачити кількість правильних і неправильних прогнозів для кожної категорії.

На основі цієї матриці будується ряд інших характеристик (accuracy, precision, recall, f1-score). Розглянемо кожну з них більш детально.

Акуратністю (англ. Accuracy) називається пропорція точних прогнозів по відношенню до загальної кількості прогнозів, тобто це ймовірність того, що клас буде передбачений правильно (1.1).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}\ , \tag{1.1}$$

де TP – це кількість прикладів, які були правильно класифіковані як позитивні, FP –це кількість прикладів, які були неправильно класифіковані як позитивні, TN –це кількість прикладів, які були правильно класифіковані як негативні, FN – це кількість прикладів, які були неправильно класифіковані як негативні.

Точністю (англ. Precision) називається частка правильних відповідей моделі в межах класу - це частка об'єктів, що дійсно належать даному класу, щодо всіх об'єктів які система віднесла до цього класу. Формула для розрахунку (1.2):

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{1.2}$$

де TP – це кількість прикладів, які були правильно класифіковані як позитивні, FP –це кількість прикладів, які були неправильно класифіковані як позитивні.

Повнота (англ. Recall) – це частка істинно позитивних класифікацій. Повнота показує, яку частку об'єктів, що реально належать до позитивного класу, ми передбачили вірно. Або ж іншими словами: це частка варіантів, класифікованих як позитивні, які насправді виявилися позитивними. Формула для розрахунку (1.3):

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{1.3}$$

де TP – це кількість прикладів, які були правильно класифіковані як позитивні, FN – це кількість прикладів, які були неправильно класифіковані як негативні.

F-міра (f1-score) є гармонійним середнім між точністю і повнотою. Вона прагне до нуля, якщо точність або повнота прагне до нуля. Формула для розрахунку (1.4):

$$\text{f1 score} = \frac{2*Precision*Recall}{Precision+Recall}, \quad (1.4)$$

де Precision та Recall визначаються за формулами (1.2) та (1.3).

1.3.2 Метод дерево рішень (Decision Tree)

Дерево рішень є потужним інструментом, використовуваним в області машинного навчання та аналізу даних. Воно дозволяє створити модель, яка може класифікувати дані та приймати рішення на основі вхідних параметрів.

Дерево рішень – це графічна модель, яка представляє рішення або послідовність рішень, засновану на вхідних параметрах. Воно складається з вузлів та гілок, де кожен вузол представляє певний тест або умову, а гілки вказують на можливі варіанти відповідей або наступні дії. “Дерева рішень” використовуються для класифікації, прогнозування та підтримки прийняття рішень.

Дерево рішень застосовується у:

- логістична регресія. “Дерева рішень” можуть використовуватись для логістичної регресії, що допомагає вирішувати задачі класифікації. Вони можуть передбачати ймовірність того, що певний об’єкт належить до певного класу, на основі набору вхідних параметрів;
- класифікація та прогнозування. “Дерева рішень” добре підходять для класифікації об’єктів на основі їх властивостей. Наприклад, вони можуть класифікувати пацієнтів на групи ризику на основі медичних показників або передбачити, чи буде клієнт відмовлятися від послуги на основі його покупок;
- підтримка прийняття рішень . “Дерева рішень” можуть допомагати в прийнятті рішень в умовах невизначеності. Вони можуть аналізувати різні вхідні параметри та рекомендувати найкращі варіанти дій на основі встановлених критеріїв.

Дерево рішень складається зі специфічної структури, яка включає вузли та гілки. Розглянемо основні елементи структури:

- вузли та гілки;

Кожен вузол в Дереві рішень представляє певний тест або умову, що має бути перевірена. Наприклад, вузол може містити питання “Чи велике число X?”. Гілки вказують на можливі відповіді або наступні дії, залежно від результату тесту. Наприклад, якщо відповідь на питання “Чи велике число X?” є “Так”, то дерево перейде до наступного вузла, якщо відповідь є “Ні”, то до іншого.

- представлення даних;

Вхідні дані, що використовуються для побудови Дерева рішень, можуть бути представлені у вигляді таблиці або матриці. Кожний рядок таблиці представляє окремий об’єкт, а кожний стовпець вказує на певну властивість цих об’єктів. Наприклад, вхідні параметри можуть включати вік, стать, дохід, освіту тощо.

- критерії розгалуження

Критерії розгалуження визначають, які тести або умови слід використовувати для прийняття рішень при побудові Дерева рішень. Наприклад, критерієм може бути ентропія або чистота вузлів, які оцінюються для вибору найкращого розгалуження. Оцінюючи критерії, можна вибрати найкраще Дерево рішень з декількох побудованих варіантів. Це може включати вибір дерева з найвищою точністю або оптимальним балансом між точністю та складністю.

Ентропія відповідає ступеню хаосу в системі. Чим вище ентропія, тим менше впорядкована система і навпаки. Інформація протилежна ентропії.
Ентропія Шеннона (Shannon) визначається для системи з N можливими станами так за формулою (1.5):

$$S = -\sum p_i \cdot log_2(p_i), \qquad (1.5)$$

де p_i - частота входження елемента (наприклад, ймовірність того, що випадково обраний фільм буде саме цього жанру). У свою чергу, вона обчислюється за такою формулою (1.6):

$$p_i = \frac{N_i}{N}, \tag{1.6}$$

де N - загальна кількість елементів ; N_i - кількість елементів певного.

Основою алгоритмів побудови дерева рішень є принцип жадібної максимізації приросту інформації – на кожному кроці вибирається та ознака, за якою під час розподілу приріст інформації виявляється найбільшим. Далі процедура повторюється рекурсивно, поки ентропія не буде дорівнювати нулю або якійсь малій величині (якщо дерево не підлаштовується ідеально під навчальну вибірку, щоб уникнути перенавчання). У різних алгоритмах застосовуються різні евристики для «ранньої зупинки» або «відсікання», щоб уникнути побудови перенавченого дерева. [12]

Дерева рішень мають свої переваги та недоліки.

Переваги:

- простота інтерпретації: Дерева рішень легко інтерпретувати, оскільки вони можуть бути представлені у формі графіку або діаграми;
- використання для навчання з наглядом: Дерева рішень можуть працювати з наборами даних, які містять як категоріальні, так і числові атрибути;
- робота з великими наборами даних: Дерева рішень можуть ефективно працювати з великими наборами даних та великою кількістю атрибутів.

Недоліки:

- надмірна чутливість до шуму: Дерева рішень можуть бути чутливі до шуму або незначних змін у вхідних даних, що може призвести до неправильних рішень;
- накопичення помилок: При рекурсивному розгалуженні Дерева рішень можуть накопичувати помилки, оскільки кожне розгалуження базується на попередніх рішеннях;

- недостатня універсальність: Дерева рішень можуть бути обмежені у своїй здатності моделювати складні взаємозв’язки між атрибутами. [11]

1.3.3 Метод випадковий ліс (Random Forest)

Random forest (з англ. - «випадковий ліс») - алгоритм машинного навчання, що полягає у використанні ансамблю вирішальних дерев. Алгоритм 49 поєднує в собі дві основні ідеї: метод Беггінга Бреймана, і метод випадкових підпросторів запропонований Тін Кам Хо. Алгоритм застосовується для задач класифікації, регресії і кластеризації. Основна ідея полягає в використанні великого ансамблю вирішальних дерев, кожне з яких саме по собі дає дуже невисоку якість класифікації, але за рахунок їх великої кількості результат виходить хорошим.

RF (random forest) - це безліч вирішальних дерев. У задачі регресії їх відповіді усереднюються, в завданні класифікації приймається рішення голосуванням за більшістю. Всі дерева будуються незалежно за наступною схемою:

- Вибирається підвибірка навчальної вибірки розміру – по ній будується дерево (для кожного дерева - своя підвибірка)
- Для побудови кожного розщеплення в дереві переглядаємо max_features випадкових ознак (для кожного нового розщеплення – свої випадкові ознаки)
- Вибираємо найкращу ознаку і розщеплення по ній (за заздалегідь заданим критерієм). Дерево будується, як правило, до вичерпання вибірки (поки в листі не залишаться представники тільки одного класу), але в сучасних реалізаціях є параметри, які обмежують висоту дерева, число об'єктів в листі і число об'єктів в підвибірці, при якому проводиться розщеплення. [14]

Зрозуміло, що така схема побудови відповідає головному принципу ансамблювання (побудови алгоритму машинного навчання на базі кількох, в даному випадку вирішальних дерев): базові алгоритми повинні бути хорошими

і різноманітними (тому кожне дерево будується на своїй навчальній вибірці і при виборі розщеплення є елемент випадковості).

Чим більше дерев, тим краща якість, але час налаштування і роботи RF також пропорційно збільшуються. Часто при збільшенні кількості дерев якість 50 на навчальній вибірці підвищується (може навіть доходити до 100%), а якість на тестовій вибірці виходить на асимптоту.

Принцип роботи наведено на рисунку 1.3

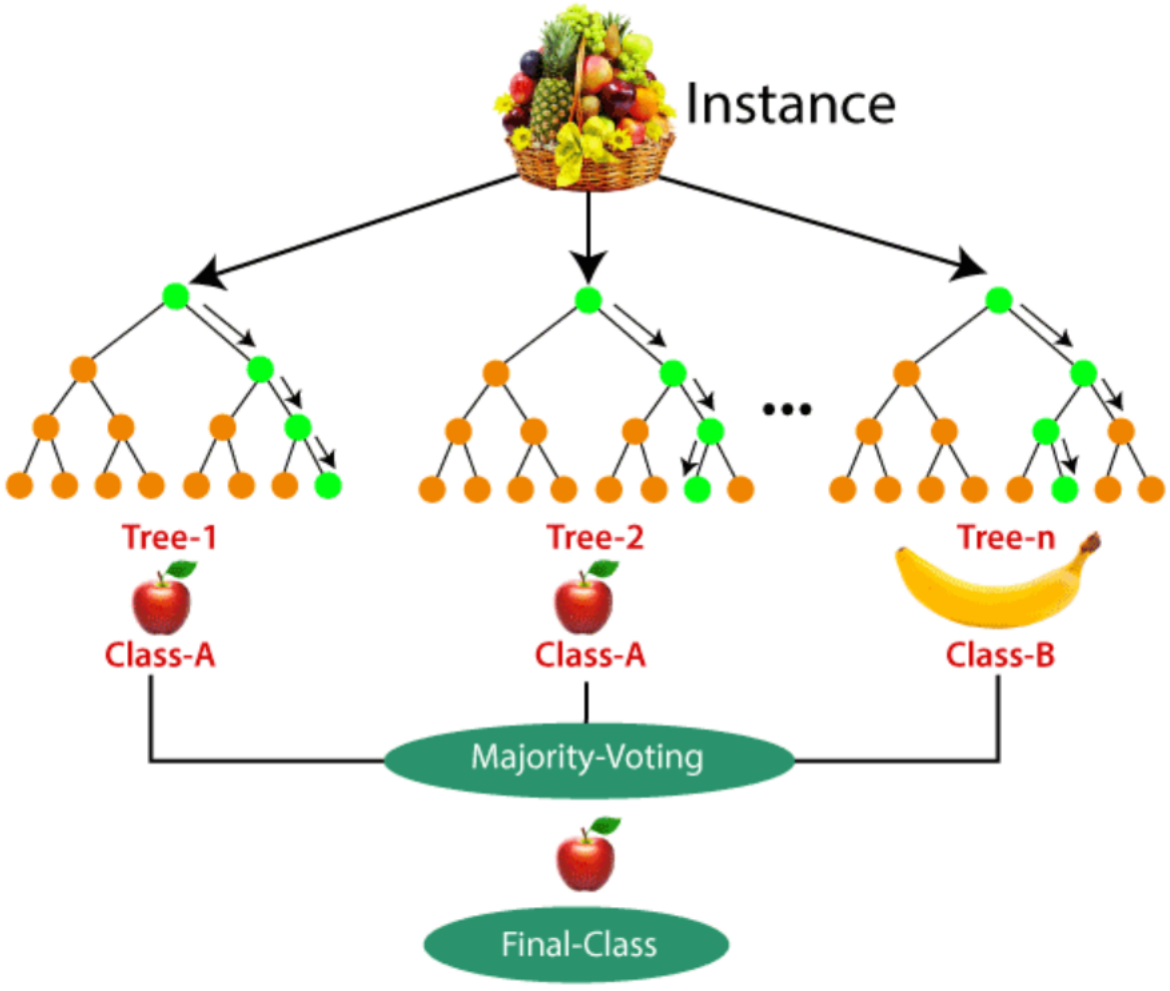


Рисунок 3.3 – Принцип роботи алгоритму Random Forest

Переваги Random Forest:

- висока точність: Random Forest зазвичай має високу точність прогнозів, оскільки він комбінує результати багатьох рішучих дерев, що дозволяє зменшити вплив випадкових помилок та перенавчання;

- стійкість до перенавчання: Благодаря випадковому вибору підмножини даних та ознак на кожному кроці, Random Forest є менш схильним до перенавчання порівняно з окремим рішучим деревом;
- взаємодія з багатьма типами даних: Він може ефективно використовувати як числові, так і категоріальні ознаки, а також працювати з даними різних типів, такими як текст, зображення тощо;
- здатність ранжування ознак: Random Forest може автоматично оцінити важливість ознак в задачі, дозволяючи визначити, які ознаки мають найбільший вплив на результати;
- зменшення дисперсії: Випадковий ліс може допомогти зменшити дисперсію прогнозів в порівнянні з окремим рішучим деревом.

Недоліки Random Forest:

- складність моделі: Random Forest може бути весьма складним за рахунок великої кількості рішучих дерев, що в ньому використовуються, що може збільшувати час навчання та обчислювальні витрати;
- важкість інтерпретації: Через свою складність, Random Forest може бути важким для інтерпретації. Визначення важливості окремих ознак може бути нетривіальним завданням;
- потреба в підборі гіперпараметрів: Для досягнення найкращих результатів може знадобитися налаштування гіперпараметрів, таких як кількість дерев у лісі, глибина дерев і інші;
- неефективність для рідкісних класів: Якщо у вас є дисбаланс класів у даних і один клас є дуже рідкісним, Random Forest може бути менш ефективним для прогнозування цього класу.

Загалом, Random Forest є потужним і дуже корисним алгоритмом для багатьох задач машинного навчання. Однак важливо враховувати його обмеження та докладати зусиль для оптимізації та встановлення правильних гіперпараметрів для вашого конкретного завдання. Також цей метод краще використовувати для задач класифікації.[15]

1.3.4 Метод k-найближчих сусідів (K-NN)

K-Nearest Neighbour (K-NN) – один із найпростіших алгоритмів машинного навчання, що базується на методі контрольованого навчання.

Алгоритм K-NN передбачає схожість між новим випадком/даними та доступними спостереженнями та поміщає новий випадок у категорію, найбільш схожу на доступні категорії.

Алгоритм K-NN зберігає всі доступні дані та класифікує нову точку даних на основі подібності. Це означає, що з появою нових даних їх можна легко класифікувати за категоріями свердловин за допомогою алгоритму K-NN.

Алгоритм K-NN може бути використаний як регресії, так класифікації, але в основному він використовується для завдань класифікації.

K-NN є непараметричним алгоритмом, що означає, що він робить ніяких припущень з урахуванням базових даних.

Його також називають алгоритмом відкладеного навчання, тому що він не навчається з навчального набору одразу, а зберігає набір даних та під час класифікації виконує дію з набором даних.

Алгоритм KNN на етапі навчання просто зберігає набір даних, а коли він отримує нові дані, він класифікує ці дані за категоріями, які схожі на нові дані.

Роботу K-NN можна пояснити на основі наступного алгоритму:

Крок-1: Виберіть кількість K сусідів

Крок-2: Обчислити евклідову відстань K числа сусідів

Крок-3: Візьмемо K найближчих сусідів відповідно до обчисленої евклідової відстані.

Крок-4: Серед цих k сусідів підрахуйте кількість точок даних у кожній категорії.

Крок 5: Визначте нові точки даних тієї категорії, для якої число сусідів є максимальним.

Крок-6 Наша модель готова. [16]

Принцип роботи алгоритму зображено на рисунку 1.4.

Переваги k-NN:

- простота реалізації: k-NN - один з найпростіших алгоритмів машинного навчання, і він легко реалізовується;
- не вимагає побудови моделі: k-NN не будує явну модель, тобто він не вчиться з даних. Це означає, що він може бути використаний в ситуаціях, де дані не мають якоїсь чіткої структури або закономірностей;
- здатність робити класифікацію для складних, нелінійних задач: k-NN може

добре справлятися з завданнями, в яких класи не розділені лінійно;

- діє для задач класифікації та регресії: Інша перевага полягає в тому, що k-NN може бути використаний як для задач класифікації, так і для задач регресії, просто змінюючи спосіб обчислення вихідного значення.

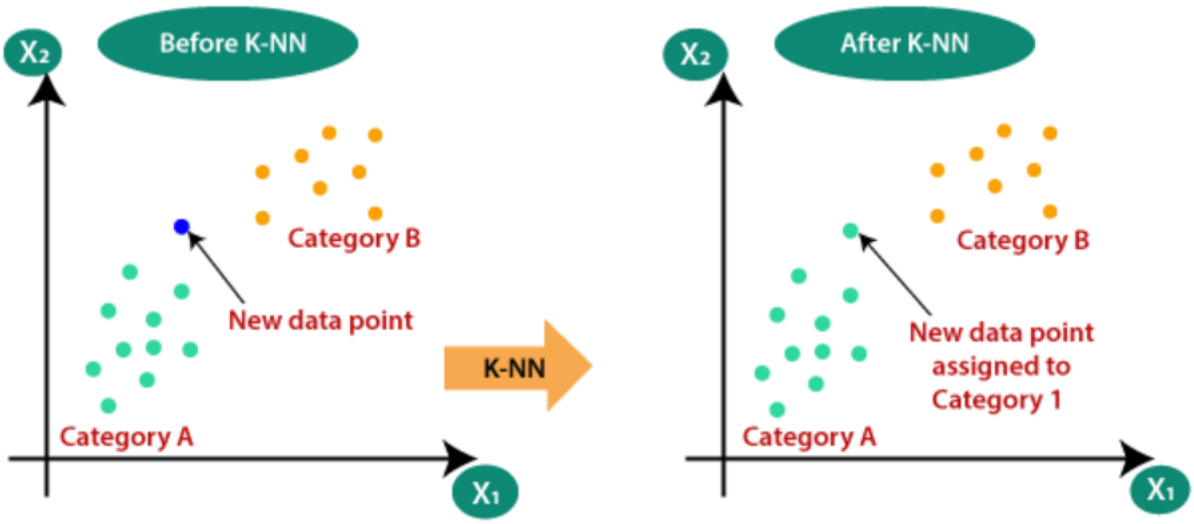


Рисунок 4.4 – Принцип роботи алгоритму K-NN

Недоліки k-NN:

- обчислювальна складність: Один із головних недоліків k-NN - це висока обчислювальна складність, особливо при великому обсязі даних і великому значенні k. Розрахунок відстаней між всіма точками може бути витратним у великих вибірках;
- чутливість до шуму: k-NN може бути чутливим до шуму і викидів в даних.

Один викид може суттєво вплинути на результати класифікації або регресії.

- неявна обробка категоріальних ознак: k-NN не добре обробляє категоріальні ознаки, і для їх використання може знадобитися додаткова попередня обробка даних.
- Потребує налагодження параметру k: Визначення оптимального значення k (кількості сусідів) може бути нетривіальним завданням. Занадто мале значення k може призвести до перенавчання, а занадто велике - до згладжування класів.

Усі ці фактори роблять k-NN потужним інструментом, але з обмеженнями. Цей алгоритм добре підходить для випадків, де дані мають просту структуру та не мають великої обсягової складності. [17]

1.3.5 Метод Ada Boost

Ada Boost (Adaptive Boosting) - це алгоритм адаптивного збільшення, який використовується для покращення точності моделей машинного навчання, особливо слабких моделей (наприклад, слабких класифікаторів). Цей алгоритм включає в себе декілька базових моделей і вивчає їх послідовно, надаючи більший вага тим прикладам, які були неправильно класифіковані попередніми моделями. [18]

Основні концепції та принципи роботи алгоритму Ada Boost включають наступне:

- початкова вага для прикладів: Початково всі приклади мають однакову вагу;
- побудова базової моделі (слабкого класифікатора): Перша базова модель навчається на навчальних даних з вагами прикладів. Мета - знайти модель, яка класифікує приклади найкраще можливим чином;
- оцінка помилок: Після побудови першої моделі обчислюється загальна помилка на навчальних даних, де вага кожного прикладу залежить від того, чи був він правильно класифікований;

- вага моделі: Обчислюється вага для моделі на основі її точності. Моделі, які правильно класифікували багато прикладів, отримують більшу вагу;
- оновлення ваг прикладів: Ваги прикладів змінюються таким чином, щоб підкреслити приклади, які були неправильно класифіковані попередньою моделлю;
- побудова наступної базової моделі: Наступна базова модель навчається на даних з оновленими вагами прикладів. Процес побудови і оновлення моделей повторюється декілька разів;

- ансамблювання моделей: Після побудови всіх базових моделей, їх результати комбінуються з вагами, враховуючи їхню точність;
- кінцевий класифікатор: Отриманий ансамбль моделей стає кінцевим класифікатором. Приклад класифікується на основі голосування моделей - більше ваги дається думці більш точних моделей. [18]

Переваги методу AdaBoost включають покращену точність класифікації, високу стійкість до перенавчання і здатність працювати добре з різними базовими моделями. Однак, важливо враховувати, що AdaBoost може бути вразливим до шуму в даних, і невірна настройка параметрів може призвести до перенавчання.

1.3.6 Аналіз наукових статей (досліджень), які були проведені раніше

Зараз цукровим діабетом хворіють дуже багато людей у різних країнах. Найбільш поширений є в Індії, США, Китаю, Мексиці та у країнах південно-східної Азії. На разі Україна належить до середнього рівня поширення діабету, але з кожним років відсоток захворюваності на цю хворобу, на жаль, збільшується. Тому дана робота полягає у можливості виявлення цукрового діабету до його появи аби запобігти його розвитку в організмі людини, використовуючи показники та звички людини за допомогою методів машинного навчання.

Насамперед були розглянути наукові статі та дослідження, які вже були проведені на цю тему.

Аналіз літератури на дану тему було проведено у науковій базі даних Google Scholar на англійській та українських мовах. На даний запит українською мовою було знайдемо 342 статті, але на подібні теми, де був суто аналіз виникнення діабету, прогнозування діабету з використанням нейронних мереж та інші. Результат зображено на рисунку 1.5.

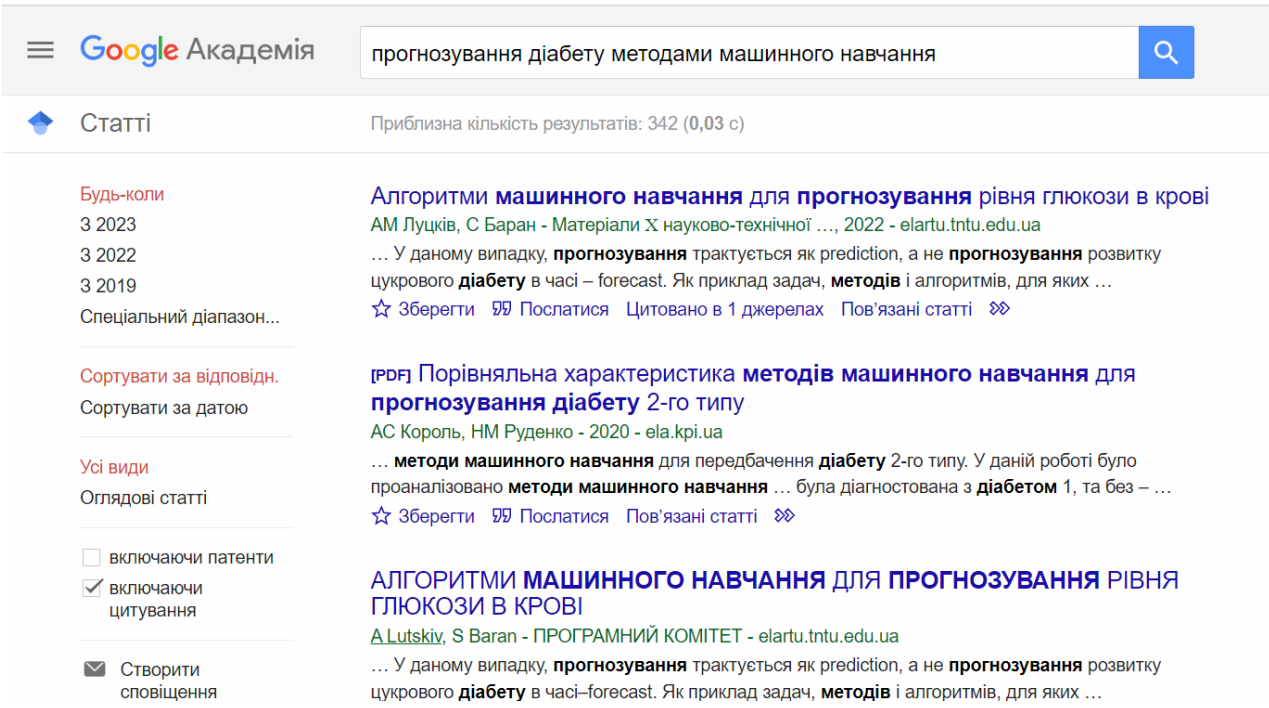


Рисунок 5.5 – Результати пошуку наукових досліджень українською мовою

Проте на англійській мові на запит прогнозування цукрового діабету методами машинного навчання “machine learning diabetes prediction” було знайдено 23500 статей, де у деяких статтях використовували інші технології для прогнозування чи хоч одне слово назви статті збіглося з нашим запитом. Тому далі були обрані точні критерії для відбору літератури, де вже було знайдено 30 наукових статей. Результат наведено на рисунках 1.6 та 1.7.

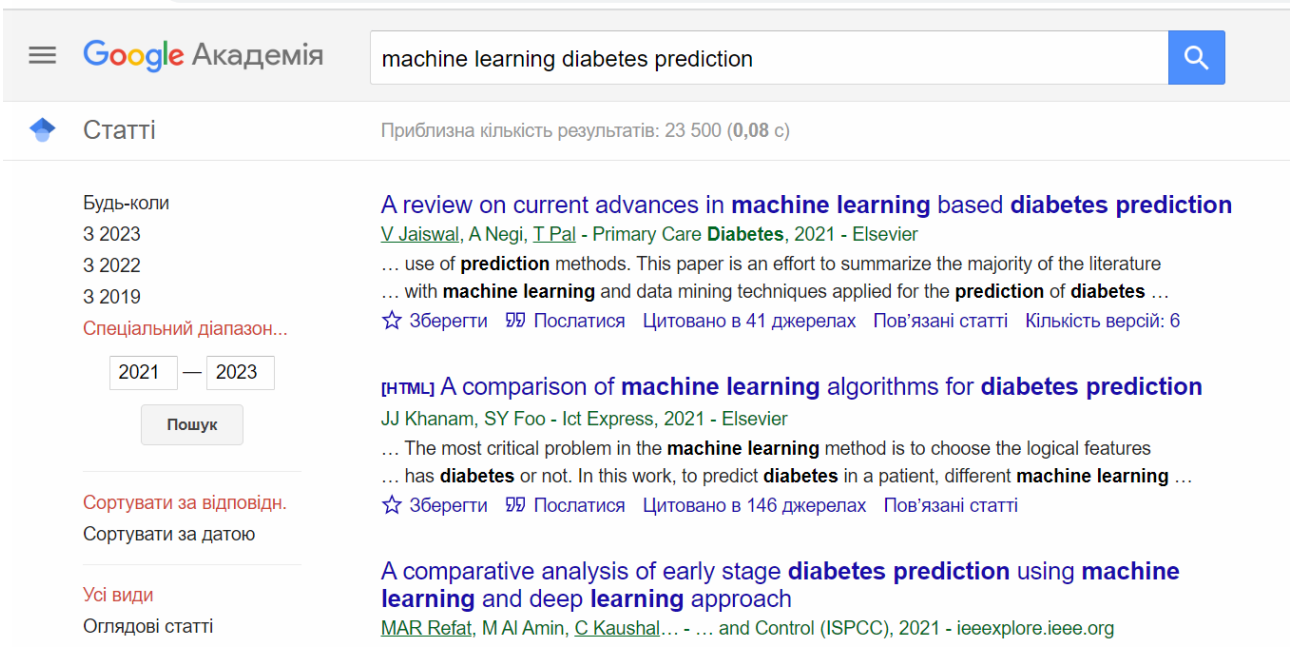


Рисунок 6.6 – Результати пошуку наукових досліджень англійською мовою

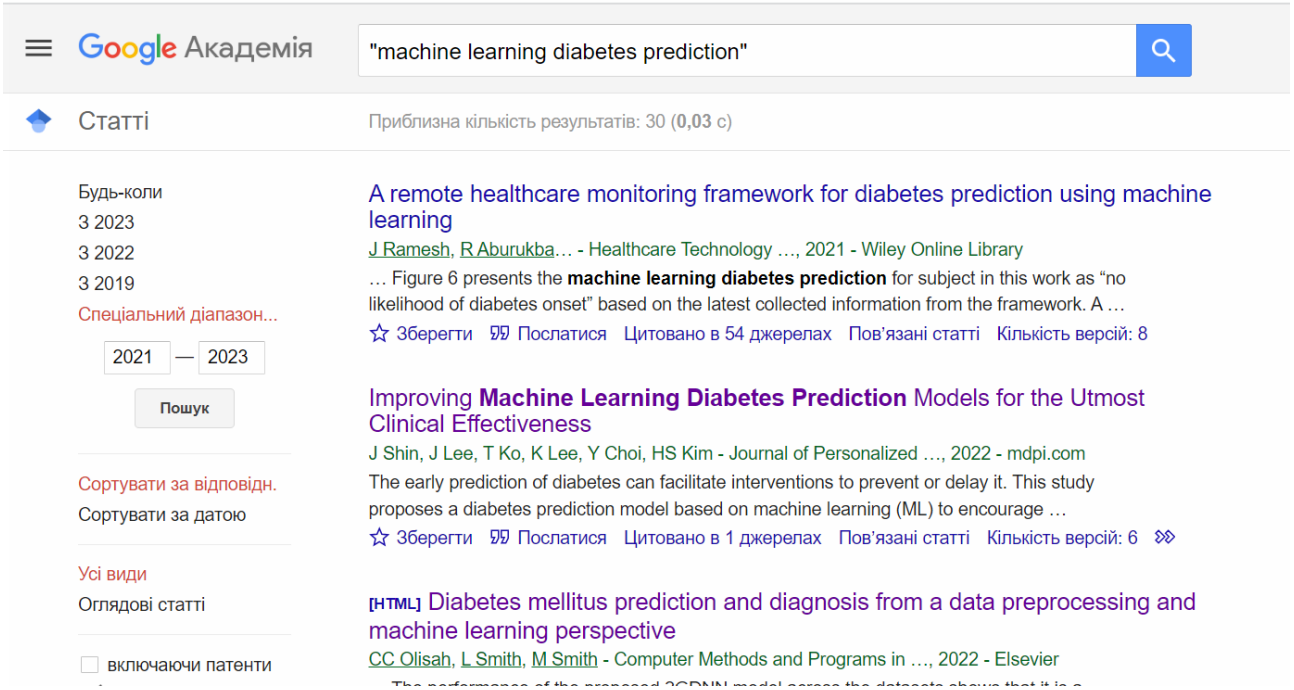


Рисунок 7.7 – Результати пошуку наукових досліджень англійською мовою з певними критеріями пошуку

Проаналізувавши дані статті, виявилося, що:

- деякі з них не доступні для загального перегляну;
- доступна тільки анотація;
- дослідження проведено для дуже малої кількості, що не є релевантним для

аналізу та проведення висновків.

Тож було обрано та проаналізовано три статті, які більш за всі пов’язані з даною темою кваліфікаційної роботи.

У статті [19] автори провели дослідження, яке пропонує підхід для запобігання появи діабету. Була створена модель прогнозування діабету методами машинного навчання, а саме: Decision tree, Random forest, XGBoost, Cox regression. Результати показали, що найбільшу точність отримала модель 81,2% за допомогою використання методу XGBoost. Але є і недоліки даного дослідження, що були використані обмежені дані (які були взяті з одної клініки), не були враховані фактори способу ведення життя та поганих звичок.

У статі [20] автори провели дослідження, яке пропонує підхід для запобігання появи діабету. Була створена модель прогнозування діабету методами машинного навчання, а саме: Random forest (RF) model, Support vector machine (SVM) model, twice-growth deep neural network (2GDNN). Результати показали, що найбільшу точність отримала модель 97,3% за допомогою методу 2GDNN та 92,26% - методу Random Forest.

Але є і недоліки даного дослідження, що не були враховані фактори способу ведення життя, поганих звичок, статі людини та те, що отримані найкращі результати були для власного метода авторів, який не є доступний для кожної людини.

У статті [21] автори провели дослідження, яке пропонує підхід для запобігання появи діабету. Була створена модель прогнозування цукрового діабету методом машинного навчання Random forest (RF). Результати показали, що модель отримала точність 88,14%.

Але є і недоліки даного дослідження, що не були використані інші методи для проведення дослідження.

Наразі вже є багато досліджень методами машинного навчання, які використовуються для оцінки ймовірності того, що людина захворіє на діабет. Але деякі з досліджень проводилися суто для одної групи/ клініки пацієнтів або не враховували всі фактори, що можуть вплинути на хворобу. Тому для подальшого припинення розвитку чисельності хворих на цукровий діабет

потрібно далі проводити аналіз та дослідження факторів, які можуть вплинути на діабет й прогнозування виникнення цукрового діабету у здорової людини за певними факторами.

1.4 Висновку до розділу

У інформаційно-аналітичному розділі було розглянуто основні теоретичні відомості про штучний інтелект, зокрема машинного навчання. Ціль машинного навчання – частково або й повністю автоматизувати рішення різних складних аналітичних задач.

Далі було розглянуто його головні аспекти: основні етапи машинного навчання, складові машинного навчання (дані, ознаки та алгоритми), методи класифікації (дерево рішень, випадковий ліс, k-найближчих сусідів та AdaBoost). Також було проведено аналіз критеріїв (accuracy, presicion, recall, f1-score та матриця помилок) для оцінювання якості моделі класифікації, а саме для визначення точності моделі при прогнозуванні.

Наостанок було проаналізовано наукові дослідження та статті за допомогою Google Scholar, які є вже на просторах інтернету з теми прогнозування виявлення цукрового діабету у людини методами машинного навчання.

Тому, можна зробити висновок, що машинне навчання покликане давати максимально точні прогнози на підставі вступних даних, щоб лікарі могли завчасно виявляти цукровий діабет у людини до появи його симптомів чи дуже стрімкого підняття рівня глюкози у крові людини.

2 СПЕЦІАЛЬНИЙ РОЗДІЛ

2.1 Постановка задачі

Цукровий діабет є однією з найбільш поширеніших захворювань у світі.

Сьогодні близько 530 мільйонів людей хворію на цукровий діабет у світі я, серед них 1 300 000 людей – українці станом на червень 2023 року. Діабет - це хронічне захворювання, яке впливає на рівень цукру (глюкози) в крові.

Важливо зауважити, що діабет – це серйозна проблема громадського здоров'я, і він може призвести до різних ускладнень, якщо не контролюється належним чином. . Поява діабету зазвичай є результатом поєднання генетичних, середовищних та стилевих факторів. Ось деякі з головних факторів, які спричиняють появу діабету: генетична схильність, ожиріння, неправильна харчова поведінка, інсулін резистентність та шкідливі звички людини.

Одним з головних факторів, які можуть сприяти виникнення цукрового діабету є відсутність своєчасного виявлення захворювання у людини. Зазвичай, люди, які мають такі симптоми: сухість у роті, часті сечовипускання, погіршення зору, постійну втому, підвищення чи втрату ваги, постійне відчуття голоду не звертаються до лікаря. Бо вони не знають, що дані симптоми можуть вказувати на високий рівень глюкози у крові.

Тому, задача даної кваліфікаційної роботи полягає в тому, що потрібно проаналізувати причини та зробити прогнозування виявлення цукрового діабету до його появи аби запобігти його розвитку в організмі людини, використовуючи показники та звички людини. Оскільки, діабет може призвести до серйозних ускладнень, включаючи пошкодження нирок та серцево-судинних захворювань, що загрожує життю. У разі недоуправління цим захворюванням можливі важкі стани, такі як ампутація кінцівок та втрата зору.

Тоді задача полягає у:

- аналізі факторів, які впливають на появу цукрового діабету у людини;
- прогнозуванні виявлення цукрового діабету за допомогою методів машинного навчання (Decision tree, Random Forest, k-NN, Ada Boost);
- дослідженні ефективність цих методів машинного навчання для знаходження кращого методу для розв'язання даної задачі.

2.2 Аналіз початкових даних

Початкові дані для аналізу та прогнозування виявлення цукрового діабету до його появи аби запобігти його розвитку в організмі людини було взято з сайту «Центри контролю та профілактики захворювань» [22]. Дані було подано у форматі файлу *csv, який складаються з 22 колонок (показники та параметри людей) та 253689 рядків (значення показників та параметрів) для кожної людини.

Записи початкового файлу наведено у додатку Г.

Структура даних наведено у таблиці 2.1.

Таблиця 2.1

Структура даних

Назва колонки	Тип (значення) колонки	Опис колонки
-1-	**-2-**	**-3-**
Diabetes_012	Число (0,1,2) 0 – no diabetes 1 – prediabetes 2 – diabetes	Інформація, який описує 0 – пацієнт не має діабету, 1 – пацієнт схильний до діабету (перед діабет), 2 – пацієнт має діабет
HighBP	Число (0,1) 0 – no high BP 1 – high BP	Інформація чи має високий тиск пацієнт: 0 – не має високий тиск, 1 – має високий тиск.
HighChol	Число (0,1) 0 – no high cholesterol 1 – high cholesterol	Інформація про пацієнта чи має високий холестерин: 0 – не має високий холестерин, 1 – має високий холестерин.

CholCheck	Число (0,1) 0 – no 1 – yes	Інформація про пацієнти про перевірку холестерину через 5 років:0 – не було,1 – було

Продовження табл. 2.1

-1-	-2-	-3-
BMI	Число	Індекс мали тіла
Smoker	Число (0,1) 0 – no 1 – yes	Інформація про пацієнти, чи курив як мінімум 100 цигарок за своє життя: 0 – ні, 1 – так.
Stroke	Число (0,1) 0 – no 1 – yes	Інформація про пацієнта, чи був колись інсульт: 0 – ні, 1 – так.
HeartDiseaseorAttack	Число (0,1) 0 – no 1 – yes	Інформація про пацієнта, чи була колись ішемічна хвороба серця або інфаркт: 0 – ні, 1 – так.
PhysActivity	Число (0,1) 0 – no 1 – yes	Інформація про фізичну активність пацієнту протягом останні 30 днів, окрім роботи: 0 – ні, 1 – так.
Fruits	Число (0,1) 0 – no 1 – yes	Інформація про споживання фруктів 1 або більше разів у день пацієнта: 0 – ні, 1 – так.
Veggies	Число (0,1) 0 – no 1 – yes	Інформація про споживання овочів 1 або більше разів у день пацієнта: 0 – ні, 1 – так.
HvyAlcoholConsump	Число (0,1) 0 – no 1 – yes	Інформація про споживання алкоголю (для чоловіків більше 14 напоїв на тиждень, для жінок – 7): 0 – ні, 1 – так.
AnyHealthcare	Число (0,1) 0 – no 1 – yes	Інформація про наявність медичного страхування у пацієнта: 0 – ні, 1 – так.
NoDocbcCost	Число (0,1) 0 – no 1 – yes	Інформація про похід пацієнта до будь-якого лікаря протягом останніх 12 місяців: 0 – ні, 1 – так.
GenHlth	Число (1-5) 1 – excellent, 2 – very good, 3 – good, 4 – fair, 5 – poor.	Інформація про здоров’я пацієнту на його думку за шкалою: 1 – відмінна, 2 – дуже хороша, 3 – нормальна, 4 – задовільна, 5 – погана.

Продовження табл. 2.1

-1-	-2-	-3-
MenthHlth	Число (1-30)	Інформація про психічне здоров'я, де пацієнт мав стрес, депресію та проблеми з емоціями протягом останніх 30днів. Треба записати скільки днів пацієнт мав такий стан від 1 до 30.
PhysHlth	Число (1-30)	Інформація про фізичне здоров'я, де пацієнт мав фізичні захворювання та травми протягом останніх 30днів. Треба записати скільки днів пацієнт мав такий стан від 1 до 30.
DiffWalk	Число (0,1) 0 – no 1 – yes	Інформація про пацієнта чи є серйозні труднощі при ходьбі чи підйомі по сходах: 0 – ні, 1 – так.
Sex	Число (0,1) 0 – female 1 - male	Стать пацієнта: 0 – жінка, 1 – чоловік.
Age	Число	Вік, де поділено на 13 категорій: 1 – 18-24, 13 – 80 і більше років.
Education	Число (1-6) 1 – never attended school 2 – elementary grades 3 – some high school graduate 4 – high school graduate 5 – college year to 3 years 6 – college 4 years or more	Інформація про освіту пацієнта: 1 – ніколи не відвідував школу, 2 – 1-8 класи, 3 – 9-11 класи, 4 – 12 класів, 5 – коледж до 3 років, 6 – закінчив коледж.
Income	Число (1-8)	Інформація про заробітну плату, яка розподілена на 8 груп, де 1 – менше ніж 10000$, 8 – більше, ніж 75000$.

Для подальшого аналізу та прогнозування методами машинного навчання мовою програмування Python даний дата сет було завантажено у програмне середовище Colab для подальшої обробки даних. На рисунках 2.1 – 2.2

продемонстровано початковий дата сет у програмному середовищі.

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
253675	0.0	1.0	1.0	1.0	45.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0
253676	2.0	1.0	1.0	1.0	18.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	4.0	0.0
253677	0.0	0.0	0.0	1.0	28.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0
253678	0.0	1.0	0.0	1.0	23.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0
253679	2.0	1.0	1.0	1.0	25.0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	2.0	0.0

253680 rows × 22 columns

Рисунок 2.1 – Вигляд перших 5 та останніх 5 рядків даних дата сету у програмному середовищі Colab

I	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0
0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0
0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	5.0	0.0	1.0	5.0	6.0	7.0
0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	4.0	0.0	0.0	1.0	0.0	11.0	2.0	4.0
0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	5.0	2.0
0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	0.0	0.0	1.0	7.0	5.0	1.0
0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	9.0	6.0	2.0

Рисунок 2.2 – Вигляд перших 5 та останніх 5 рядків даних дата сету у програмному середовищі Colab

Набір даних у даному дата сеті є незбалансованим, оскільки значення 0 (не має діабету) зустрічається 213703 рази, 1 (перед діабет) зустрічається 35346 раз та 2 (має діабет) зустрічається 4631 раз. Результат наведено на рисунку 2.3. Цей фактор потрібно врахувати в процесі попереднього опрацювання даних

```
0.0    213703
2.0     35346
1.0      4631
Name: Diabetes_012, dtype: int64
```

Рисунок 2.3 – Розподіл початкових даних на класи по кількості початкових значень

2.3 Попереднє опрацювання даних

Попереднє опрацювання даних є важливою складовою у багатьох сферах життя, де інформація потребує обробки, аналізу та використання для прийняття рішень. Для даної кваліфікаційної роботи є важливою сфера медицини, де дані пацієнтів обробляються для поліпшення діагностики та лікування пацієнтів.

Процедура попереднього опрацювання даних є важливим етапом у методах машинного навчання і містить ряд кроків, спрямованих на підготовку та очищення даних перед їх поданням моделям машинного навчання.

Оскільки, реальних наборів даних для машинного навчання дуже сприйнятливі до відсутності, непослідовності та шуму через їх неоднорідне походження.

Застосування алгоритмів інтелектуального аналізу даних до цих шумних даних не дасть якісних результатів, оскільки вони не зможуть ефективно ідентифікувати закономірності. Тому обробка даних є важливою для покращення загальної якості даних.

Основні кроки попереднього опрацювання даних включають наступне:

– збір та завантаження даних: Спочатку потрібно зібрати або завантажити дані з відповідних джерел, таких як бази даних, файли CSV, API тощо, наведено на рисунках 2.1 та 2.2.

– очищення даних (Data Cleaning): Визначення та видалення відсутніх значень (NaN або NULL), дублікатів та інших помилок у даних. Спочатку було видалено дані, які мали дублікати, кількість знайдених дублікатів зображено на рисунку 2.4.

```
data.duplicated().sum()
```

```
15548
```

Рисунок 2.4 – Кількість дублікатів, які були знайдені у дата сеті

Після видалення дублікатів даний дата сет містить171622 рядків. Значень NaN або NULL не було знайдено у даному дата сеті, результат наведено на рисунку 2.5.

```
data.isnull().sum().any()

False
```

Рисунок 2.5 – Перевірка даних дата сету на наявність значень NaN або NULL

– візуалізація даних (Data Visualization): Аналіз та візуалізація даних, щоб отримати більше інсайтів про їх розподіл та взаємозв'язки між ознаками.

Оскільки, початковий набір даних був поданий закодовано по категоріальних ознаках, тому за допомогою методу «мапування». Він полягає в тому, що ви створюєте відповідність між числовими значеннями й текстовими категоріями. Ви визначаєте, які числові значення будуть замінені якими категоріями.

– нормалізація та стандартизація (Normalization and Standardization): Приведення числових ознак до одного і того ж масштабу або стандартної одиниці вимірювання.

– видалення викидів (outliers) - це важлива складова попереднього опрацювання даних, і це важливо зробити, коли у ваших даних є значення, які суттєво відрізняються від інших і можуть спотворити результати аналізу або моделі. Викиди можуть бути результатом помилок вимірювань, випадкових аномалій або відображати реальні несподівані явища. [24]

Для даного дата сету було здійснення викидів поступово. Тобто спочатку було видалено аномальні значення для колонки індекс маси тіла (BMI) тому, що то була єдина колонка, яка мала значення, які не були поділені на категорії, зображено на рисунку 2.6.

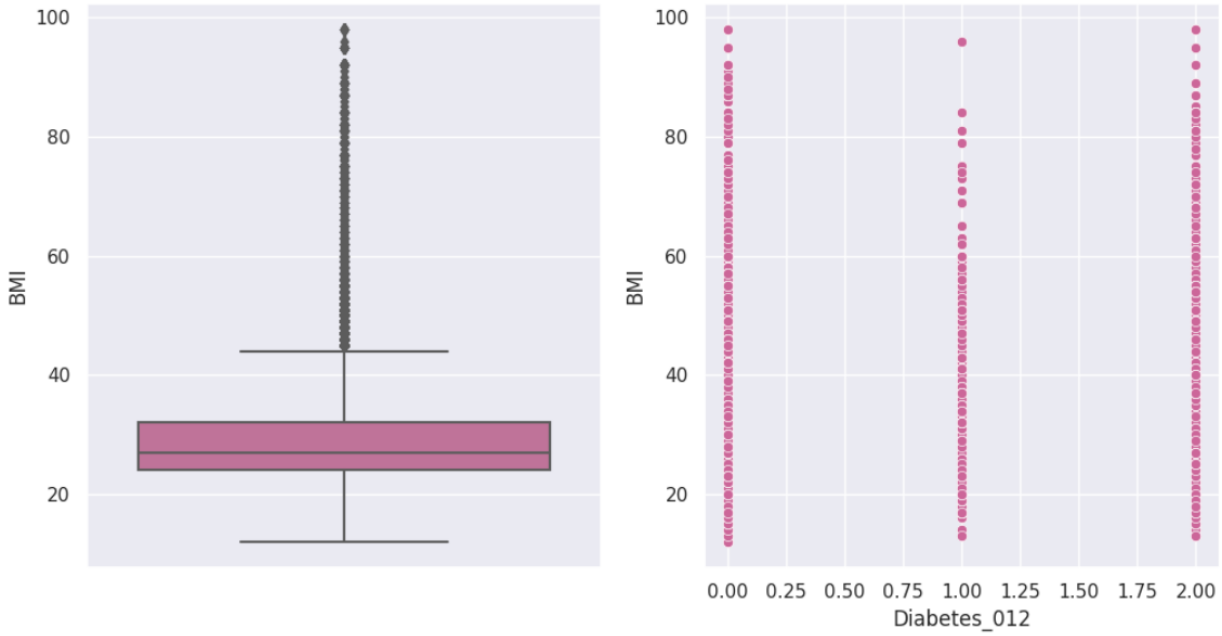


Рисунок 2.6 – Коробковий та точковий графіки викидів фічі індекс маки тіла (BMI) до видалення аномалій

З даної колонки були видалені всі значення, які перевищують 70, так як це аномалія. Результат наведено на рисунку 2.7.

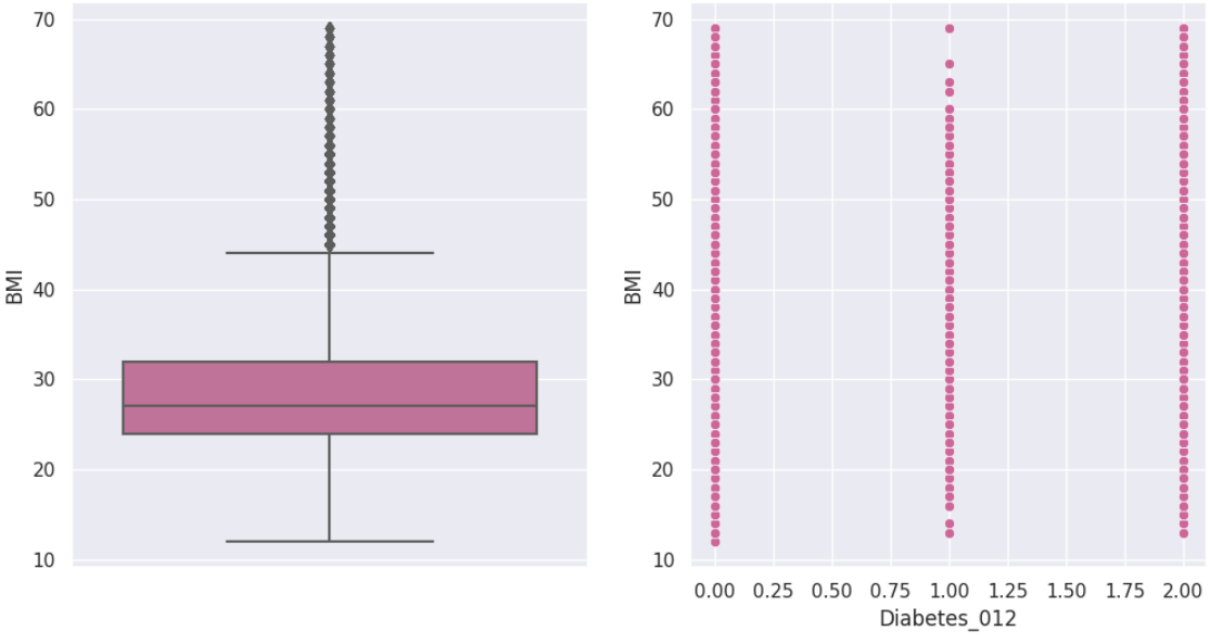


Рисунок 2.7 – Коробковий та точковий графіки викидів фічі індекс маки тіла (BMI) після видалення аномалій

Далі було виконано видалення викидів за допомогою методу Isolation Forest.

Isolation Forest — це неконтрольований алгоритм машинного навчання для виявлення аномалій. Як випливає з назви, ізольований ліс — це ансамблевий метод (подібний до випадкового лісу). Іншими словами, він використовує середнє значення прогнозів за кількома деревами рішень, коли призначає остаточну оцінку аномалії даній точці даних. На відміну від інших алгоритмів виявлення аномалій, які спочатку визначають, що є «нормальним», а потім повідомляють про все інше як аномальне, Isolation Forest намагається ізолювати аномальні точки даних із самого початку. [25]

За допомогою даного методу було знайдено та видалено 114071 аномальних значень, тобто після видалення аномалій дата сет має 114071 рядків значень показників пацієнтів. Результат наведено на рисунку 2.8.

```
#Drop column(anomaly)
df.drop(columns=['anomaly'], inplace=True)
df.shape
```

```
(114071, 22)
```

Рисунок 2.8 – Результат видалення аномалій з набору даних

– кодування категоріальних ознак (Encoding Categorical Features) – це процес

перетворення якісних (категоріальних) ознак або змінних в числовий формат, який може бути використаний в алгоритмах машинного навчання. Категоріальні ознаки є незалежними від порядку та не мають внутрішнього числового значення. Вони можуть бути, наприклад, кольорами, марками автомобілів, країнами або категоріями товарів.[26]

Існують різні методи кодування категоріальних ознак, серед яких найпоширеніші наступні: one-hot або label encoding.

Для даного набору даних не було зроблено кодування категоріальних ознак тому, що дані спочатку вже було подані закодовано, зображено на рисунках 2.1 та 2.2.

– відбір ознак (Feature Selection) – це процес вибору підмножини ознак зі всього набору даних з метою покращення продуктивності моделі машинного навчання. Правильний вибір ознак може покращити якість моделі, зменшити перенавчання, збільшити швидкість навчання та поліпшити зрозуміння важливих залежностей в даних. Головною метою відбору ознак є видалення зайвих, неінформативних або корельованих ознак, які можуть призвести до перенавчання або складності моделі. Вибір правильних ознак допомагає покращити загальну ефективність моделі.

Для даного набору даних при першому аналізі кожної колонки даних було виявлено, що колонки Education, Income, AnyHealthcare, є неінформативними для виявлення цукрового діабету у людини.

– розділення даних (Data Splitting): Розділення даних - це процес розподілу набору даних на дві частини: навчальну вибірку і тестову вибірку. Цей процес є важливим етапом в побудові та оцінці моделей машинного навчання. Основна мета розділення даних полягає в тому, щоб навчати модель на одній частині даних і оцінювати її продуктивність на іншій, щоб визначити, наскільки добре модель генералізує свої знання на нові дані.

Навчальна вибірка – це підмножина даних, яку використовують для навчання моделі. Модель адаптується до цих даних, вивчаючи внутрішні залежності та патерни.

Тестова вибірка – це інша підмножина даних, яку використовують для оцінки продуктивності моделі. Тестова вибірка не була використана під час навчання і служить для перевірки, наскільки добре модель генералізує свої знання на нові дані.

Для даного набору даних було обрано, що 70% даних – навчальна вибірка та 30% – тестова.

– балансування класів (Class Balancing) – це процес управління нерівноважністю класів у наборі даних, особливо в задачах класифікації. Коли кількість прикладів одного класу набагато більша або менша, ніж кількість прикладів іншого класу, це може впливати на продуктивність моделі машинного навчання. Нерівноважність класів може призвести до перенавчання на більш численний клас, внаслідок чого модель може показувати погану продуктивність на менш численному класі.

Для даного набору даних було обрано метод SMOTEENN.

SMOTEENN (Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors) - це комбінований метод балансування класів, який поєднує два різні підходи для роботи з нерівноваженими класами в задачах класифікації. SMOTEENN поєднує SMOTE, що використовує синтетичкі екземпляри для збільшення представництва менш численного класу, і ENN, який використовує техніку видалення надлишкових прикладів з більш численного класу.

Ось як працює метод SMOTEENN:

– SMOTE (Synthetic Minority Over-sampling Technique): Спершу використовується SMOTE для створення синтетичких прикладів для менш численного класу. SMOTE вибирає кожен приклад з менш численного класу та генерує нові синтетичні приклади, додавши їх між ним і його найближчими сусідами.

– ENN (Edited Nearest Neighbors): Після використання SMOTE застосовується ENN. ENN аналізує всі приклади в наборі даних, визначає, які з них є надлишковими для більш численного класу, і видаляє їх.

SMOTEENN має на меті збалансувати класи, збільшуючи представництво менш численного класу за допомогою синтетичних прикладів і одночасно зменшити надлишковість більш численного класу за допомогою видалення

зайвих прикладів. Це допомагає покращити якість моделі класифікації та зменшити перенавчання, яке може виникнути при роботі з нерівноваженими даними.

SMOTEENN є одним зі способів боротьби з проблемою нерівноваженості класів та може бути корисним при роботі з задачами класифікації, де один з класів є значно менш представленим у вихідних даних.

Даний набір даних був дуже незбалансований, бо класи 0 мав 47875 значень, 1 – 790 значення та 2 – 4241. Після застосування метода SMOTEENN дані були збалансовані, де класи мали наступні значення: 0 – 27948 значення, 1 – 39597, 2 – 35628 значень.

Отже, дані тепер збалансовані та модель буде значно якісніше навчатися.

– масштабування атрибутів - це процес нормалізації значень атрибутів або ознак у наборі даних так, щоб вони мали однаковий масштаб або діапазон значень. Це важливий етап в підготовці даних для багатьох алгоритмів машинного навчання, особливо для методів, які чутливі до різниці в масштабі між атрибутами.

Для даного набору даних було обрано метод Standard Scaler.

Standard Scaler – це один з методів масштабування атрибутів, який використовується для стандартизації значень атрибутів у наборі даних. Цей метод допомагає зробити розподіл значень атрибутів більш стандартизованим, де середнє значення (середній) стає 0, а стандартне відхилення (середньоквадратичне відхилення) - 1. [28]

2.4 Обґрунтування обраних моделей машинного навчання

Сьогодні використовуються дві найбільш популярні моделі машинного навчання – це класифікація та регресія. Оскільки, задача даної кваліфікаційної

роботи полягає у прогнозування виявлення цукрового діабету у пацієнтів, де варіанти відповідей тільки три (не має діабету, перед діабет, має діабет), тобто спростування або підтвердження, тому доцільно використовувати класифікаційну модель машинного навчання, детальніше про класифікацію написано у розділі 1 пункті 3.

Тому для розв'язання поставленої задачі, яку описано у постановці задачі було використано 4 моделі класифікації: Decision Tree класифікатор, Random Forest класифікатор, K-NN класифікатор, Ada Boost класифікатор.

2.5 Процес підбору гіперпараметрів для моделей машинного навчання

Важливо вибрати правильні гіперпараметри перед початком навчання, оскільки цей тип змінної безпосередньо впливає на продуктивність кінцевої моделі машинного навчання.

У машинному навчанні модель визначається або представлена параметрами моделі. Проте процес навчання моделі передбачає вибір оптимальних гіперпараметрів, які алгоритм навчання використовуватиме для вивчення оптимальних параметрів, які правильно відображають вхідні характеристики (незалежні змінні) на мітки або цілі (залежна змінна), щоб ви досягли певної форми інтелект.

Гіперпараметри моделей машинного навчання - це налаштування, які визначають структуру та параметри моделі перед початком навчання. Вони не вивчаються моделлю під час навчання, але впливають на її поведінку та продуктивність. Гіперпараметри встановлюються користувачем або за допомогою процедур автоматичного підбору, таких як пошук по сітці або випадковий пошук.[29]

Підбір гіперпараметрів моделей машинного навчання важливий з кількох причин:

– Максимізація продуктивності моделі: Вірно налаштовані гіперпараметри дозволяють досягти максимальної продуктивності моделі. Вибір неправильних гіперпараметрів може призвести до поганої продуктивності моделі та погіршення результатів.

– Загальна генералізація: Оптимізація гіперпараметрів допомагає моделі краще генералізувати на нові дані. Правильно налаштована модель буде працювати краще на невидимих даних.

– Уникнення перенавчання: Неналежне налаштування гіперпараметрів може
призвести до перенавчання, коли модель "запам'ятовує" навчальні дані, а не вивчає загальні закономірності. Оптимізація гіперпараметрів допомагає уникнути цього явища.

– Виділення важливості гіперпараметрів: Процес підбору гіперпараметрів може допомогти визначити, які саме аспекти моделі впливають на її продуктивність, що важливо для розуміння моделі та вдосконалення її роботи.

– Ефективність ресурсів: Неправильно вибрані гіперпараметри можуть призвести до великого споживання обчислювальних ресурсів. Оптимізовані гіперпараметри можуть допомогти досягти кращої продуктивності за менше часу.[30]

Далі було проаналізовано можливі гіперпараметри для обраних моделей машинного навчання.

Decision Tree Classifier – основні гіперпараметри даної моделі є:

– критерій поділу (criterion): Визначає, який критерій використовувати для

визначення якості поділу. Зазвичай це може бути "gini" (Індекс Джині) або "entropy" (Ентропія).

– максимальна глибина дерева (max_depth): Встановлює максимальну глибину

дерева рішень. Це обмежує кількість рішень, які може приймати дерево.

– мінімальна кількість вибірок для розділення внутрішнього вузла (min_samples_split): Вказує мінімальну кількість вибірок, яка потрібна для поділу внутрішнього вузла. За замовчування приймає значення - 2.

– мінімальна кількість вибірок в листку (min_samples_leaf): Встановлює

мінімальну кількість вибірок, яка повинна бути в листку дерева. За замовчування приймає значення - 1.

– максимальна кількість ознак для поділу (max_features): Вказує максимальну

кількість ознак, які можуть бути розглянуті при кожному поділі внутрішнього вузла. [31]

Random Forest Classifier – основні гіперпараметри даної моделі є:

– кількість дерев (n_estimators): Вказує кількість дерев, які мають бути включені в ансамбль. Зазвичай більше дерев дозволяє покращити стабільність та продуктивність моделі, але при цьому збільшується час навчання. За замовченням 100 дерев.

– критерій поділу (criterion): Визначає критерій, який використовується для визначення якості поділу внутрішнього вузла в кожному дереві. Зазвичай це може бути "gini" (Індекс Джині) або "entropy" (Ентропія).

– максимальна глибина дерева (max_depth): Встановлює максимальну глибину кожного дерева в ансамблі. Це обмежує складність кожного окремого дерева.

– максимальна кількість ознак для поділу (max_features): Вказує максимальну кількість ознак, які можуть бути розглянуті при кожному поділу внутрішнього вузла в дереві. За замовчення «sqrt» - вказує, що кількість розглядуваних ознак буде рівна квадратному кореню від загальної кількості ознак в даних. Це допомагає збалансувати обрані ознаки та зменшити кореляцію між деревами в ансамблі.[32]

K-NN Classifier – основні гіперпараметри даної моделі є:

– кількість сусідів (n_neighbors): Цей параметр вказує, скільки найближчих сусідів враховувати при класифікації нового прикладу. Вибір правильної кількості сусідів важливий, оскільки від нього залежить точність та здатність до генералізації моделі. За замовченням п’ять сусідів.

– метрика відстані (metric): Вказує, яку метрику відстані використовувати для визначення відстані між прикладами. Зазвичай це Евклідова відстань (Euclidean distance), але також можна використовувати інші метрики, такі як Манхеттенська відстань (Manhattan distance) чи Косинусна відстань (Cosine distance). За замовченням – “minkowski” (відстань Мінковського).

– ваги сусідів (weights): Вказує, які ваги встановлювати для різних сусідів при класифікації. Два основних варіанти цього параметра - "uniform" (усі сусіди мають однакові ваги) та "distance" (сусіди мають ваги, обернено пропорційні відстані до них). За замовченням – "uniform".

– алгоритм обчислення найближчих сусідів (algorithm): Вказує, який алгоритм використовувати для швидкого обчислення найближчих сусідів. Зазвичай використовується "auto," який автоматично вибирає найкращий алгоритм в залежності від розміру набору даних, інші значення – “ball_tree”, “kd_tree”, “brute”. [33]

Ada Boost Classifier – основні гіперпараметри даної моделі є:

– кількість базових моделей (n_estimators): Вказує, скільки базових моделей (зазвичай це дерева рішень) будуть використані в ансамблі. Зазвичай більше

моделей дозволяє покращити продуктивність моделі, але може призвести до перенавчання. За замовченням – 50.

– тип базової моделі (base_estimator): Вказує, який алгоритм (наприклад, дерево рішень, SVM, Naive Bayes) використовувати як базову модель для навчання. Зазвичай використовується дерево рішень за замовчуванням.

– швидкість навчання (learning_rate): Визначає вагу кожної базової моделі. Мала швидкість навчання означає більше "покарання" навчальних прикладів, що допомагає уникнути перенавчання. За замовченням значення float 1.0.

Тому для моделей Decision Tree класифікатор, Random Forest класифікатор, K-NN класифікатор, Ada Boost класифікатор було використано основні гіперпараметри для роботи моделей. [34]

Висновок

У даній частині спеціального розділу для подальшого аналізу причин (факторів) та прогнозування впливу на виникнення діабету було підготовлено початкові дані, а саме:

– визначено структуру даних, початкові дані формату *.csv завантажено у програмне середовище Colab для подальшої обробки даних й визначено, що початковий набір даних є незбалансованим;

– переднє опрацьовано дані, де було зроблено очищення даних (визначення та видалення відсутніх значень), підготовлено дані для візуалізацій, видалено викиди, відібрано ознаки, розділено дані на навчальну та тестові вибірки, зроблено балансування класів та масштабування атрибутів;

– обґрунтовано вид обраних моделей машинного навчання, що для даної поставленої задачі доцільно використати класифікаційну модель машинного навчання;

– підібрано гіперпараметри для моделей машинного навчання для методів

класифікації: Decision Tree, Random Forest, K-NN, Ada Boost.

2.6 Аналіз факторів, які впливають на виявлення цукрового діабету

У даному пункті було проаналізовано фактори, які впливають на виникнення діабету для виявлення найбільш вагомих причин для зменшення кількості пацієнтів у майбутньому. Спочатку було проаналізовані початкові дані за факторами (наявність фізичної активності, ментальне здоров'я, оцінка свого здоров'я, стать, вік, індекс маси тіла, хвороби серця, інсульт, куріння, проблеми зі здоров'ям, ходьба по сходах, споживання фруктів й овочів та інші) на які відповідало близько 250 тисяч людей, тому одразу можна зробити певні висновки, що:

- лише 25 000 пацієнтів мали ішемічні хвороби серця або інфаркт;
- близько 170 000 пацієнтів мали фізичну активність кожного дня, окрім роботи;
- близько 125 000 пацієнтів кожного дня їли фрукти;
- близько 165 000 пацієнтів споживали овочі кожен день;
- більше ніж 200 000 пацієнтів не мали алкогольної залежності;
- близько 80 000 оцінили стан свого здоров'я добре та нормально;
- майже всі пацієнти мали нормальне ментальне здоров'я;
- майже 160 000 тисяч хворіли або мали травми до 5 днів протягом останнього місяця;
- близько 50 000 пацієнтів мали серйозні проблеми з підйомами по сходах та ходьбі;
- близько120 000 пацієнтів – жінки, інші чоловіки;
- найбільша кількість пацієнтів (близько 50 000) мають вік 45 – 49 років.

Далі була побудовано секторна діаграма співвідношення трьох можливих стадій діабету, де було виявлено, що 82,7% - не мають діабет, 15,3 % - мають діабет та 2% - перед діабет. Результат зображено на рисунку 2.9.

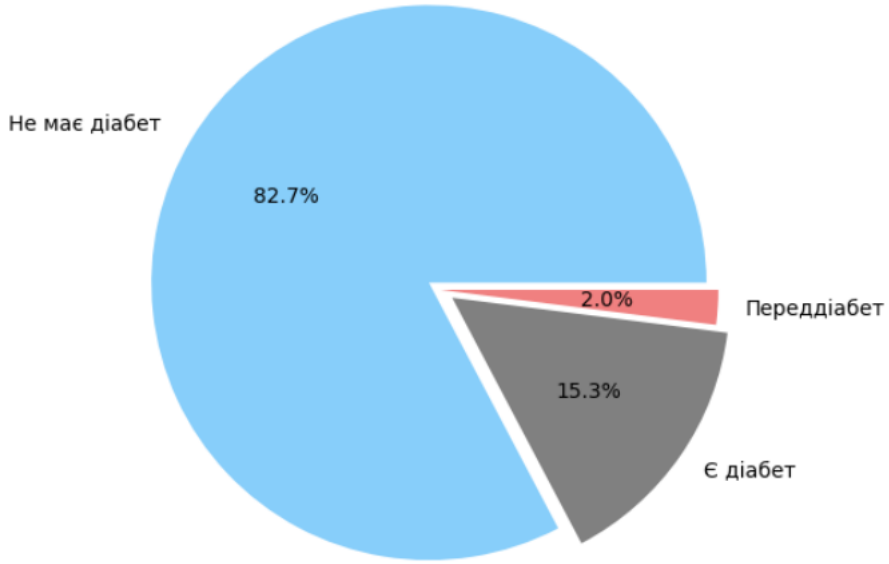


Рисунок 2.9 – Секторна діаграма співвідношення трьох можливих стадій діабету

2.6.1 Кореляція факторів, які впливають на виявлення цукрового діабету

У даному пункті було побудовано матрицю кореляції та стовпчасту діаграму кореляції для визначення факторів, які у більшому чи меншому ступені впливають на виявлення діабету. Результат зображено на рисунку 2.10 та матрицю кореляції у додатку Д.

Проаналізувавши стовпчасту діаграму та матрицю кореляції впливу факторів на цукровий діабет, можна зробити висновок, що всі фактори впливу на виявлення діабету є слабко корельовані. Але то не значить, що існує відсутність зв’язку між змінними, а вказує на то що зв’язок обмежений та має невеликий вплив.

Найбільш вплив на виявлення діабету має вплив оцінка здоров’я та артеріальний тиск – це свідчить про то, що якщо буде зростати показники високого артеріального тиску, то буде більший вплив. Гіпертензія (високий АТ) може розвиватися як результат ушкодження судин, зокрема нирок, що є спільним ускладненням діабету. Гіпертензія також може погіршити контроль рівня цукру в крові у пацієнтів з діабетом, тому цей показник є одним з найважливіших.

Кореляцію близько 0,2 мають показники холестерин, індекс маси тіла, ішемічна хвороба серця, фізичне здоров’я, труднощі з ходьбою та вік, тобто важливі фактори на вплив появи та розвитку діабету в організмі людини, а саме:

- вищий рівень холестерину у крові може збільшувати ризик розвитку діабету, оскільки дисбаланс ліпідів може впливати на інсулінову реакцію організму;
- високий індекс маси тіла (ІМТ), особливо в поєднанні з низькою фізичною активністю, є одним з основних факторів ризику для розвитку діабету, оскільки він може призводити до інсулінорезистентності;
- ішемічна хвороба серця і діабет часто взаємопоширені, оскільки обидва стани мають спільні фактори ризику, такі як високий тиск, дисліпідемія і цукровий діабет. Особи з ішемічною хворобою серця мають підвищений ризик розвитку діабету, і навпаки, діабет може погіршити перебіг ішемічної хвороби серця;
- фізичне здоров'я та труднощі з ходьбою також можуть бути пов'язані з діабетом, оскільки цей стан може призводити до ушкодження нервів та судин, що впливає на рухову активність та загальний стан здоров'я;
- вік також грає важливу роль у ризику розвитку діабету, оскільки вік є одним із найсильніших факторів ризику для розвитку діабету типу 2(який набувається за життя через неправильні харчові звички та зміни в організмі). Зі старінням зростає імовірність виникнення діабету, особливо в поєднанні з іншими факторами ризику.

Кореляцію у проміжку від 0 до 0,1 мають наступні фактори, перевірка холестерину, куріння, інсульт, наявність мед страхування, відвідування лікаря, стать та психічне здоров'я. Тобто ці фактори теж певною мірою впливають на виникнення діабету, але у значно меншому обсязі.

Від'ємну кореляцію мають фактори: фізична активність, вживання фруктів та овочів, алкогольна залежність, освіта та дохід. Тобто для даного розподілу початкового набору даних ці фактори навпаки впливають за зменшення розвитку діабету.

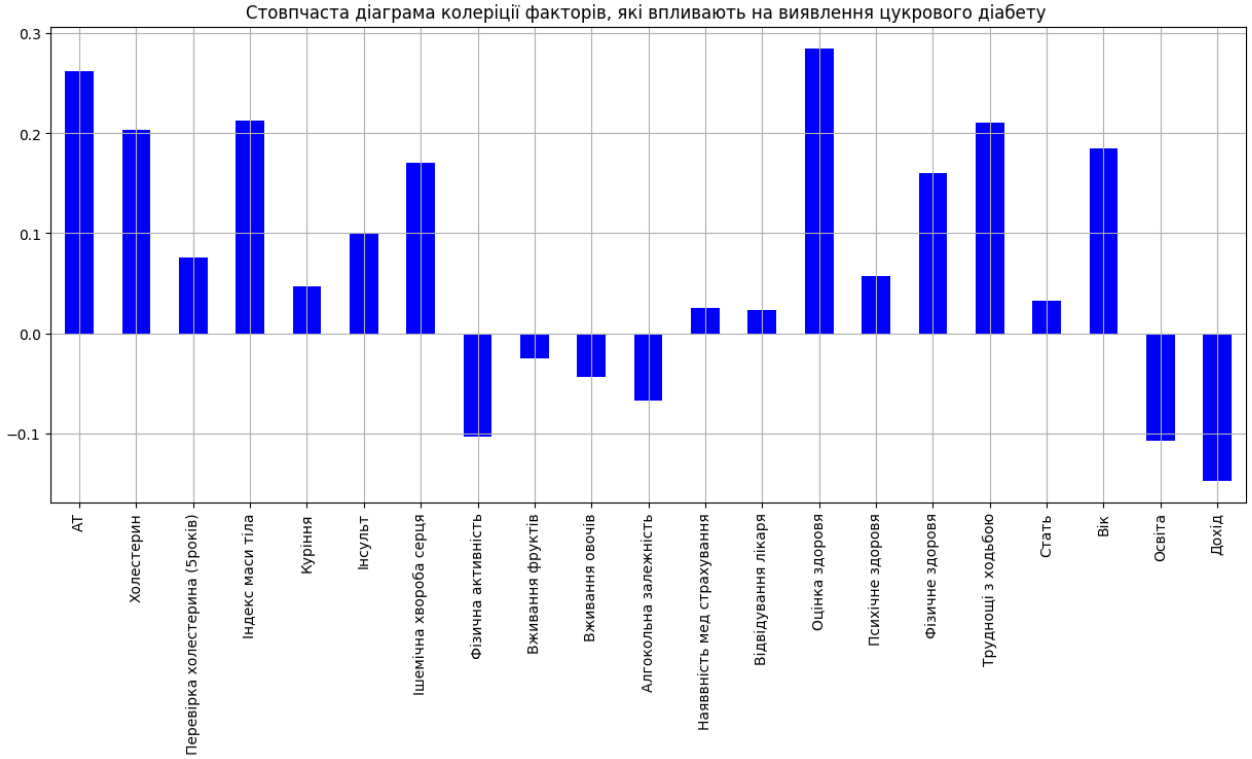


Рисунок 2.10 – Стовпчаста діаграма кореляції факторів, які впливають на виявлення діабет

Отже, можна зробити висновок, що на зменшення ймовірності розвитку цукрового діабету впливає вживання фруктів та овочів, значна більша активність, алкогольна залежність (не вживати алкоголь взагалі). А важливі фактори впливу на виявлення цукрового діабету є холестерин, індекс маси тіла, ішемічна хвороба серця, фізичне здоров'я, труднощі з ходьбою, вік, артеріальний тиск та оцінка здоров'я.

2.6.2 Графічний розподіл основних ознак в залежності від стадій діабету

У даному пункті було проаналізовано діаграми розподілу основних ознак в залежності від стадій діабету (не має діабету, є діабет, перед діабет).

На рисунку 2.11 зображено стовпчасту діаграму розподілу ознаки статі в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

- мають діабет майже однакова кількість жінок та чоловіків близько 8%;
- стан перед діабету мають лише близько 1 %, але жінки трохи більше;
- жінок – пацієнтів, які не мають діабет на 20000 більше, ніж чоловіків.

Таким чином, стан діабету або перед діабету мають трохи більше жінки, аніж чоловіки. Тому, що жінки можуть мати гормональні зміни під час життя, такі як вагітність та менопауза, які можуть впливати на ризик діабету.

Також фізична активність і харчові звички. Зазвичай більша частина жінок веде менш активний спосіб життя або має менше здорових харчових звичок, це може збільшити їхній ризик розвитку діабету.

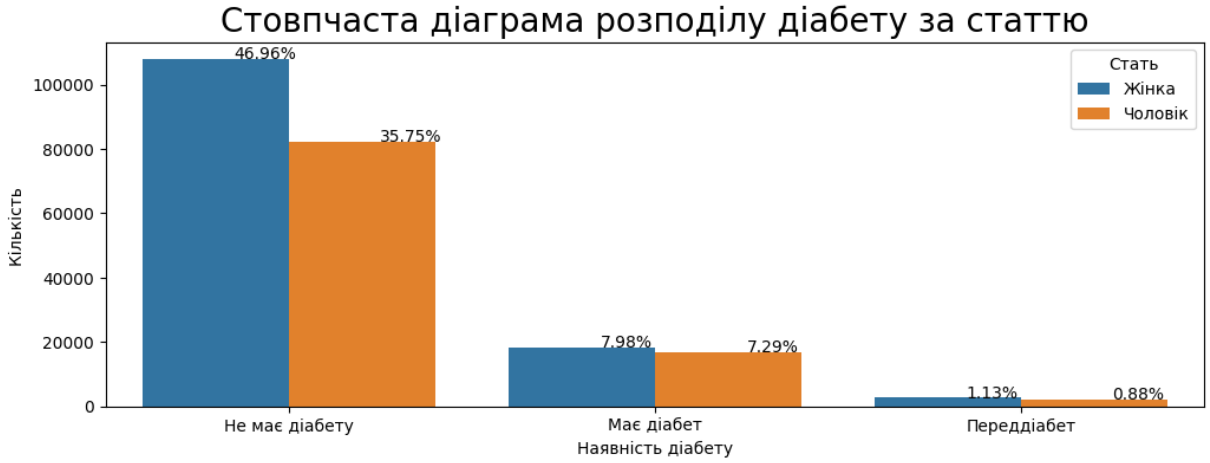


Рисунок 2.11 – Стовпчаста діаграма розподілу ознаки статі в залежності від стадій діабету

На рисунку 2.12 зображено стовпчасту діаграму розподілу ознаки куріння (тобто відповідь на запитання: чи курить пацієнт?) в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– пацієнти мають однаковий відсоток захворюваності (близько 8%) в незалежності курять вони чи ні;

– лише 27% пацієнтів курять з яких 8% мають цукровий діабет або перед діабет.

Таким чином виходить, що ознака куріння (тобто чи курить пацієнт або ні) майже не впливає на розвиток діабету в організмі людини. Але не варто забувати, що наслідки куріння, такі як набір ваги, збільшення апетиту можуть вплинути на виникнення діабету.

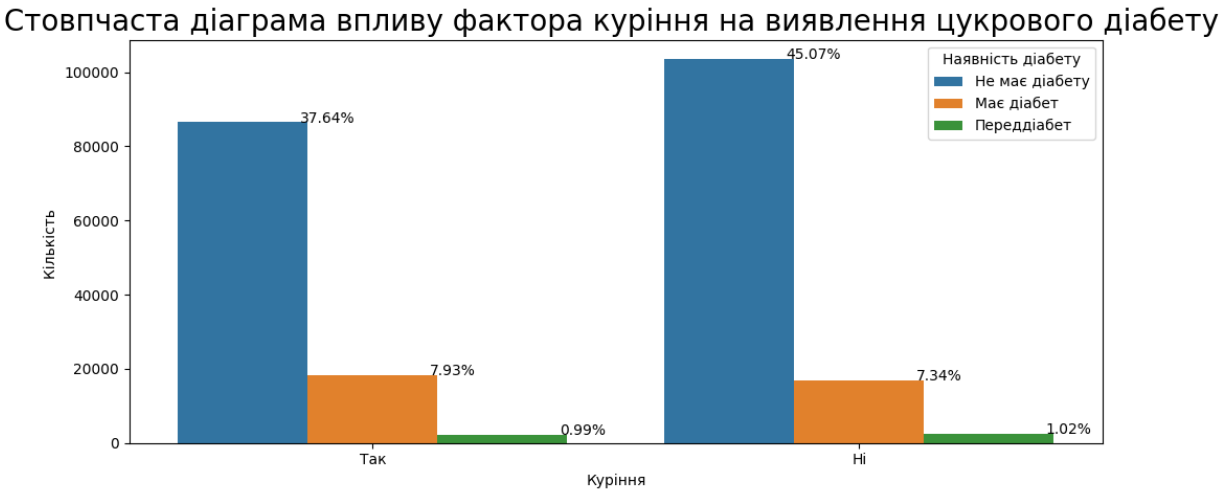


Рисунок 2.12 – Стовпчаста діаграма розподілу ознаки куріння в залежності від стадій діабету

На рисунку 2.13 зображено стовпчасту діаграму розподілу ознаки ішемічної хвороби серця в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– значна частина пацієнтів (77%), які мають ішемічну хворобу серця не мають діабету;

– частина пацієнтів, які мають ішемічну хворобу серця та діабет у 7 разів менша за пацієнтів у яких тільки ішемічна хвороба серця;

– людини, які не мають ішемічну хворобу серця хворіють на цукровий діабет

у 4 рази менше та майже у 6 разів менше мають перед діабет;

– лише 6% людей не мають ішемічної хвороби серця та цукрового діабету.

Отже, не всі пацієнти, що мають ішемічну хворобу серця можуть мати діабет та ті пацієнти, які мають ІХС хворіють приблизно у 4 рази більше на цукровий діабет. Оскільки ішемічна хвороба серця (ІХС) є серцевим захворюванням, яке виникає через обмежений кровопостачання до м'яза серця через коронарні артерії. ІСХ може впливати на розвиток діабету і наступний контроль цукру в крові у наступних способах: особливі ліки (які підіймають цукор у крові), ІХС може бути стресом для людини, що також вплине на діабет.

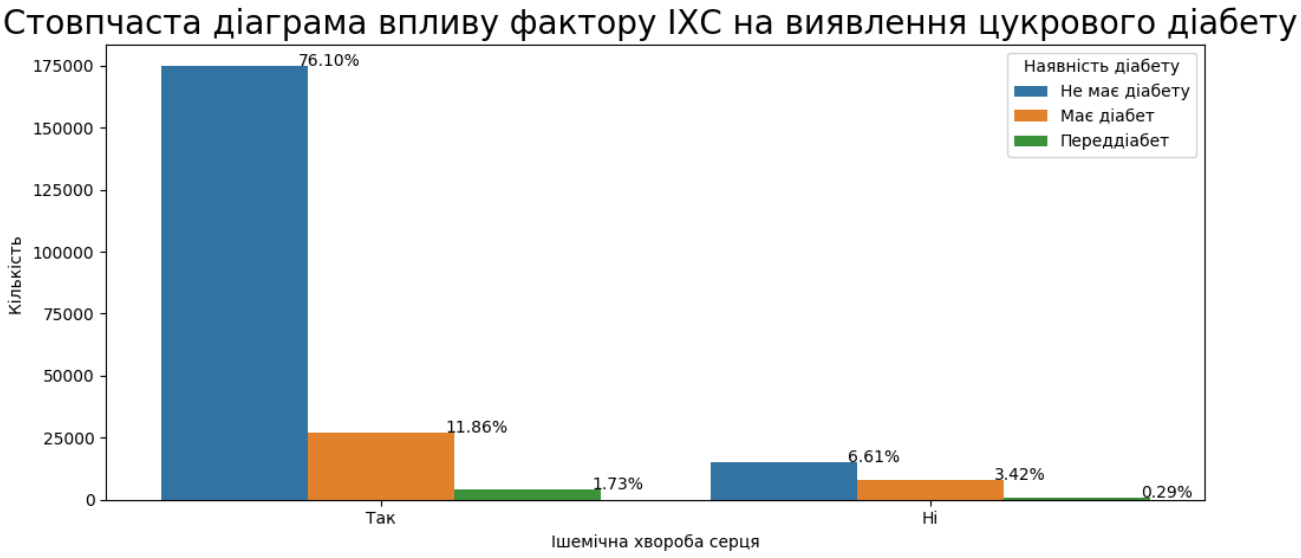


Рисунок 2.13 – Стовпчаста діаграма розподілу ознаки ішемічної хвороби серця (інфаркт) в залежності від стадій діабету

На рисунку 2.14 зображено стовпчасту діаграму розподілу ознаки вживання фруктів в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– пацієнти, які не вживають фрукти, майже у 2 рази більше не хворіють на

цукровий діабет;

– люди, які вживають фрукти, мають цукровий діабет на 1/3 частку менше, ніж інші.

Тобто, зв'язок між вживанням фруктів і ризиком розвитку цукрового діабету може бути звучним, але варто враховувати кілька важливих аспектів: вживання фруктів, зазвичай, є індикатором більш загального здорового способу життя та збалансованого харчування. Також різні фрукти мають різний вміст цукру та волокон, і вплив на рівень цукру в крові може відрізнятися від фрукта до фрукта. Також важлива кількість споживаних фруктів. Наприклад, споживання великої кількості фруктів з високим вмістом цукру може збільшити ризик розвитку діабету.

Отже, однозначного впливу споживання фруктів на виникнення цукрового діабету немає, тому що є деякі фрукти, які мають великий рівень цукру та для людей зі схильність до цукрового діабету не варто їх вживати, а є навпаки фрукти, які принесуть користь.

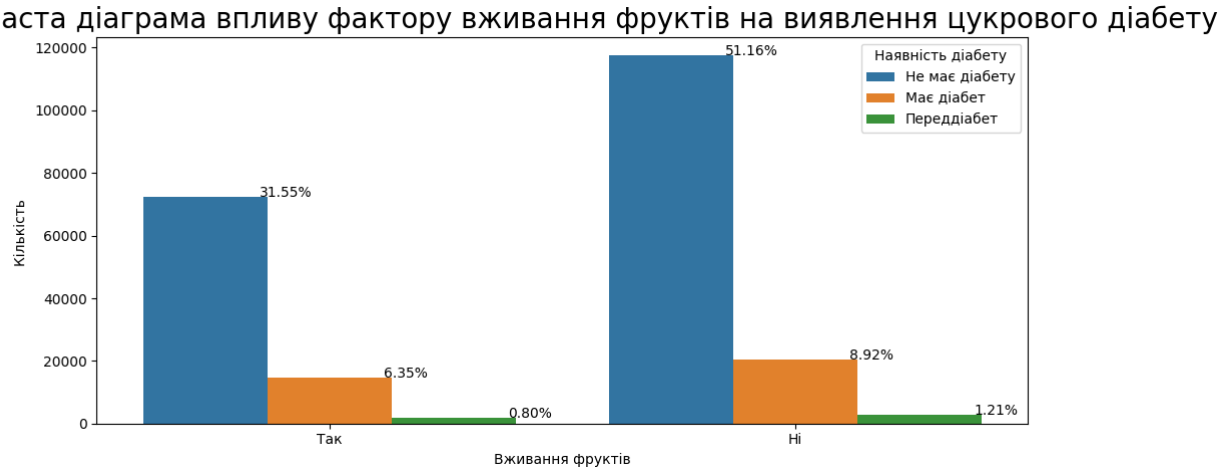


Рисунок 2.14 – Стовпчаста діаграма розподілу ознаки вживання фруктів в залежності від стадій діабету

На рисунку 2.15 зображено стовпчасту діаграму розподілу ознаки вживання овочів в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– пацієнти, які вживають овочі, майже у 6 разів більше не мають захворювання;

– люди, які споживають овочі у 4 рази більше мають діабет та у 3 рази

– перед діабет.

Спостереження, що пацієнти, які вживають овочі, мають менший ризик розвитку цукрового діабету, може бути важливим і корисним для підтримки здорового способу життя та профілактики діабету. Овочі містять важливі поживні речовини, включаючи волокна, вітаміни та антиоксиданти, які можуть впливати на ризик розвитку діабету. Але користь овочі приносять, якщо овочі якісні та дозування та баланс між всім іншим.

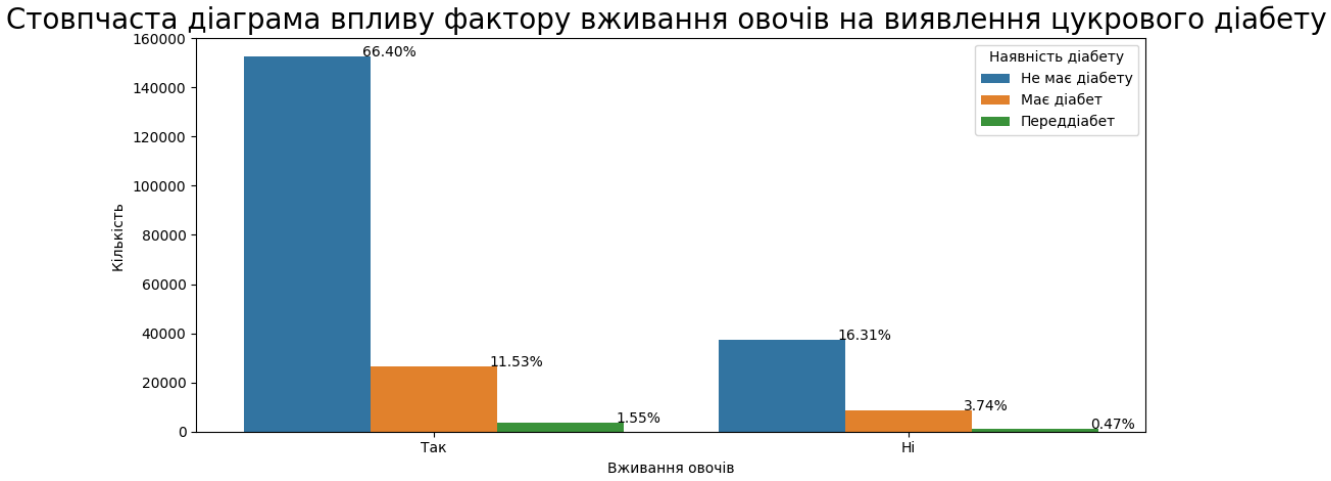


Рисунок 2.15 – Стовпчаста діаграма розподілу ознаки вживання овочів в залежності від стадій діабету

На рисунку 2.16 зображено стовпчасту діаграму розподілу ознаки алкогольної залежності в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– пацієнти у яких алкогольна залежність мають цукровий діабет у 14 разів менше, ніж люди без залежності алкоголю.

Спостереження, що пацієнти з алкогольною залежністю мають менший ризик розвитку цукрового діабету, може бути зумовлене кількома факторами, такими як припинення вживання алкоголю, можливі методологічні аспекти дослідження. Проте, важливо враховувати, що це спостереження не обов'язково вказує на причинно-наслідковий зв'язок і потребує подальших досліджень для розуміння його точних причин.

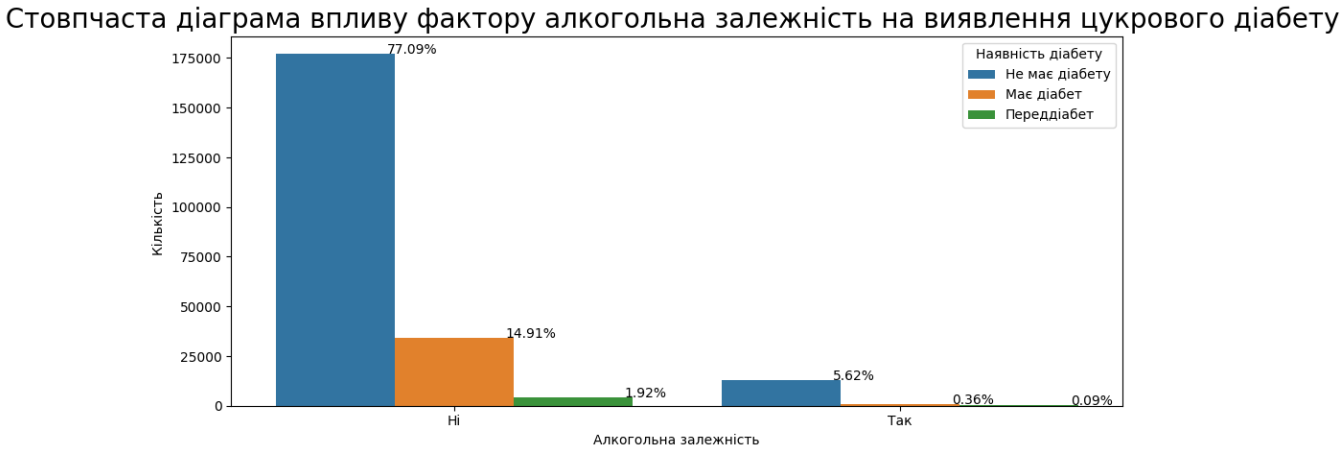


Рисунок 2.16 – Стовпчаста діаграма розподілу ознаки алкогольна залежність в залежності від стадій діабету

На рисунку 2.17 зображено стовпчасту діаграму розподілу ознаки АТ в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– кількість пацієнтів, які мають високий АТ у два рази більше, ніж з нормальний АТ при наявності цукрового діабету;

– пацієнти, які мають перед діабет приблизно однакова кількість з високим та нормальним АТ.

Отже, АТ значно впливає на виявлення цукрового діабету. Оскільки, високий артеріальний тиск (АТ) є серйозним фактором ризику для ускладнень. Він може сприяти пошкодженню судин, погіршувати контроль рівня цукру в крові й підвищувати ризик серцево-судинних захворювань та інших діабетичних ускладнень.

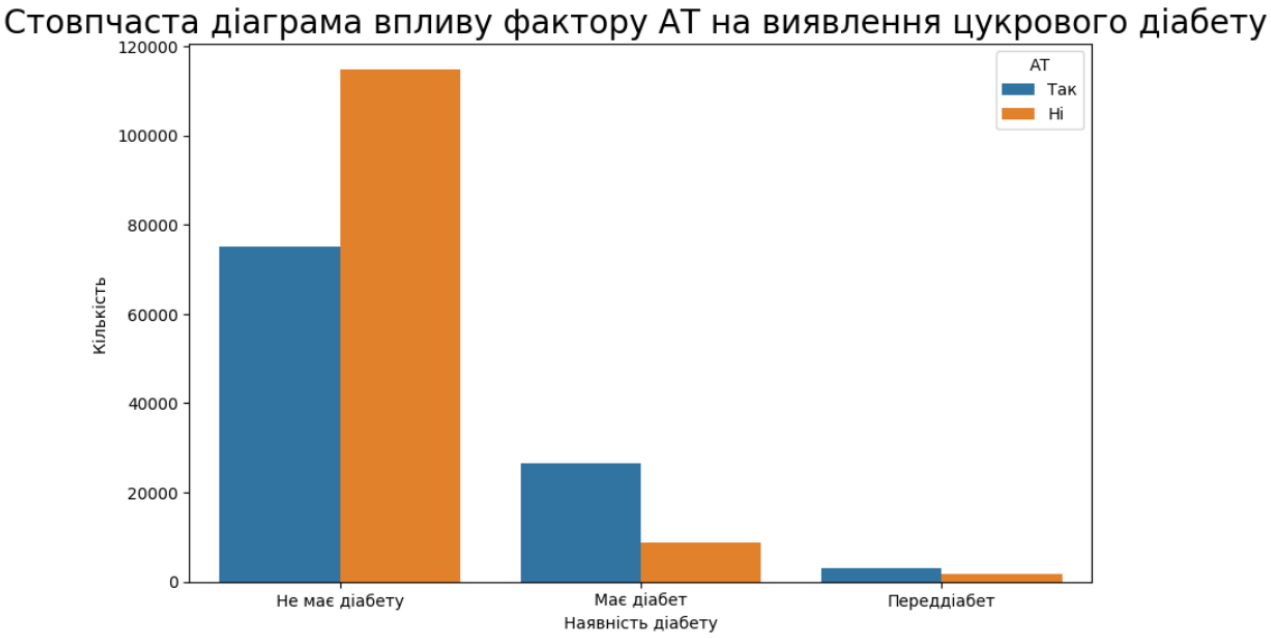


Рисунок 2.17 – Стовпчаста діаграма розподілу ознаки АТ в залежності від стадій діабету

На рисунку 2.18 зображено стовпчасту діаграму розподілу ознаки холестерину в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– пацієнти з високим холестерином мають у два рази частіше цукровий діабет;

– пацієнти, які мають перед діабет приблизно однакова кількість з високим та нормальним рівнями холестерином.

– пацієнти з нормальним рівнем холестерину на 40000 більше, які не мають діабету.

Спостереження свідчать про те, що високий рівень холестерину може бути пов'язаний зі збільшеним ризиком розвитку цукрового діабету. Однак при наявності перед діабету, рівень холестерину може бути менш важливим фактором. Пацієнти з нормальним рівнем холестерину мають менший ризик розвитку діабету, але інші фактори ризику також важливі. Важливо підкреслити, що ризик розвитку діабету є комплексним і містить багато чинників, так що необхідно брати до уваги всі аспекти здоров'я та стилі життя при оцінці ризику. Холестерин знаходиться в жирній їжі, такій як яйця, м'ясо, молоко та молочні продукти. Однак важливо пам'ятати, що організм може самостійно виробляти холестерин, тому надлишок спожитого холестерину з їжею може впливати на загальний рівень холестерину в крові.

Отже, важливо пам’ятати, що для зменшення ризику виникнення діабету потрібно вести правильний спосіб харчування, бо холестерин знаходиться у багатьох продуктах та організм виробляє його самостійно.

Рисунок 2.18 – Стовпчаста діаграма розподілу ознаки холестерину в залежності від стадій діабету

На рисунку 2.19 зображено стовпчасту діаграму розподілу ознаки ІМТ в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– більше ніж 50% пацієнтів мають ожиріння та діабет;
– надмірна вага та ожиріння пацієнтів вказують на перед діабет у близько 40% людей;
– частина людей без цукрового діабету має нормальну або надмірну вагу.

Отже, надмірна вага та ожиріння сприяють розвитку інсулінорезистентності, запаленню й іншим процесам, які можуть призвести до діабету. Тому виявлення такого зв'язку, де понад 50% пацієнтів з ожирінням мають діабет, є важливим попередженням щодо ризику та важливості здорового способу життя та контролю ваги для профілактики цукрового діабету.

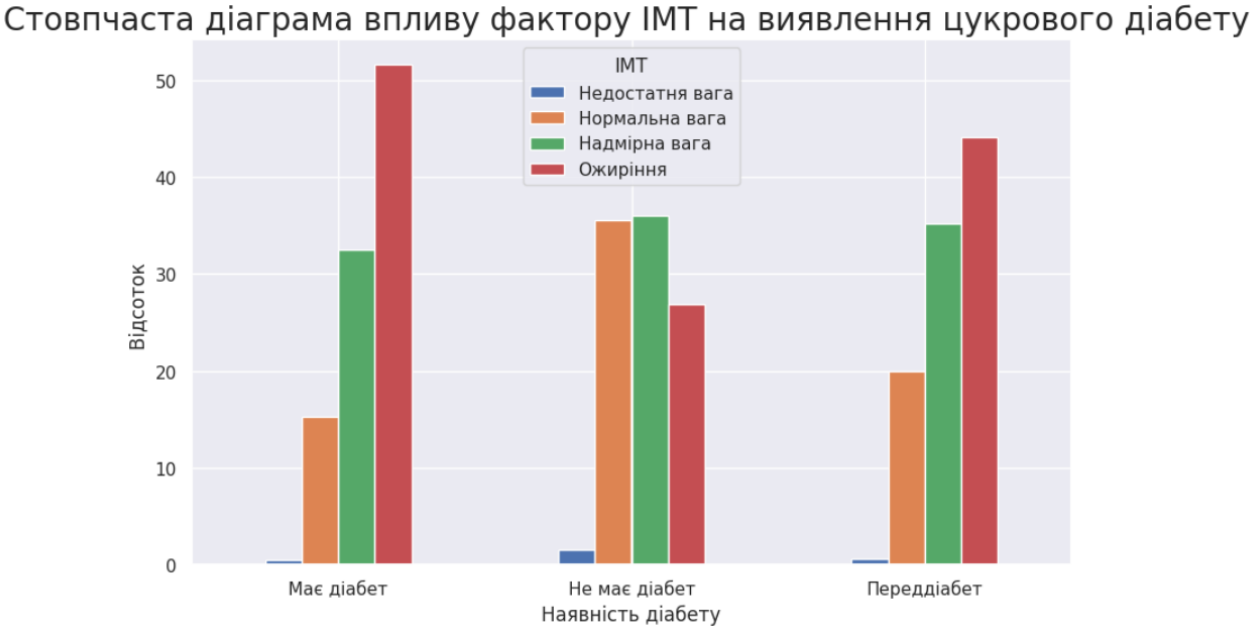


Рисунок 2.19 – Стовпчаста діаграма розподілу ознаки ІМТ в залежності від стадій діабету

На рисунку 2.20 зображено стовпчасту діаграму розподілу ознаки вік в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– найбільша кількість пацієнтів, які мають діабет у діапазоні від 50 до 64 років;

– найменша кількість пацієнтів, що мають діабет у категорії від 18 до 34 років.

В похилому віці ризик розвитку діабету збільшується через нормальні процеси старіння, такі як зниження функції підшлункової залози та збільшення інсулінорезистентності.

Загальний висновок полягає в тому, що вік грає важливу роль у ризику розвитку діабету. З віком ризик розвитку діабету типу 2 зазвичай збільшується, але цей ризик можна зменшити шляхом здорового способу життя, включаючи правильне харчування та фізичну активність. Тому важливо дотримуватися здорового способу життя та регулярно перевіряти рівень цукру в крові, особливо при зростанні віку.

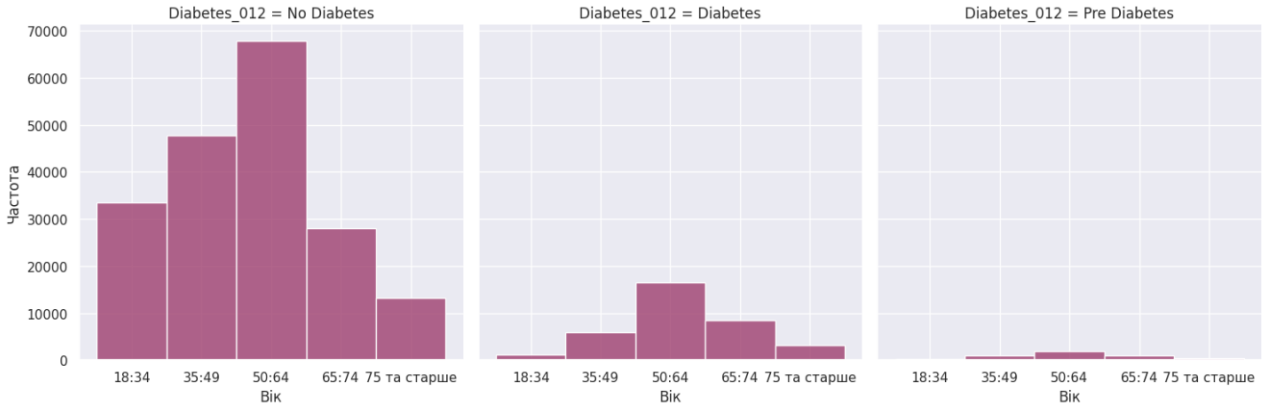


Рисунок 2.20 – Стовпчаста діаграма розподілу ознаки вік в залежності від стадій діабету

На рисунку 2.21 зображено стовпчасту діаграму розподілу ознаки оцінки здоров’я в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

– пацієнти у більшості випадків не мають діабет, якщо мають оцінку здоров’я: задовільна, нормальна, дуже хороша та відмінна;

– близько 30000 пацієнтів, що мають оцінку задовільно, нормально та дуже хороша хворіють на діабет;

– перед діабет мають пацієнти майже з усіма оцінками здоров’я.

Це може свідчити про те, що наявність діабету не завжди пов'язана з загальним станом здоров'я.

Важливо пам'ятати, що діабет може розвиватися у різних людей незалежно від їх загального стану здоров'я або оцінки його. Важливими факторами ризику розвитку діабету є генетика, стиль життя, споживання їжі та багато інших факторів.

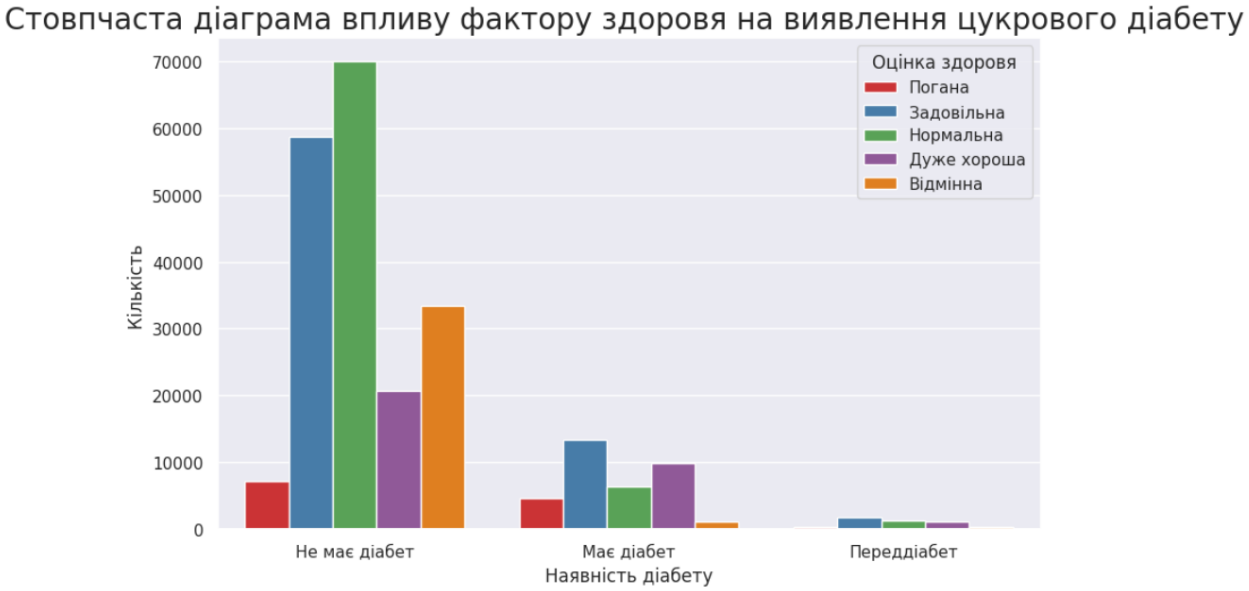


Рисунок 2.21 – Стовпчаста діаграма розподілу ознаки оцінки здоров’я в залежності від стадій діабету

Цей висновок підкреслює важливість регулярних медичних перевірок та своєчасної діагностики діабету навіть у тих людей, які можуть почувати себе добре та не мають суттєвих проблем зі своїм станом здоров'я.

На рисунку 2.22 зображено стовпчасту діаграму розподілу ознаки інсульту в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

- фактор інсульту не впливає на виникнення цукрового діабету;
- 80% пацієнтів не хворіють на діабет на не мають інсульту;
- частина пацієнтів, які мають інсульт та діабет складає лише 1,5%.

Той факт, що лише 1,5% пацієнтів мають як інсульт, так і діабет, може свідчити про те, що ці два стани не дуже часто співвідносяться один з одним

серед даної популяції. Тобто, наявність інсульту не є дуже сильним фактором ризику для розвитку діабету в цій групі пацієнтів. Варто враховувати, що ця інформація може бути корисною для оцінки співвідношення між цими двома станами та подальшого дослідження їх взаємозв'язку.

Рисунок 2.22 – Стовпчаста діаграма розподілу ознаки інсульту в залежності від стадій діабету

На рисунку 2.23 зображено стовпчасту діаграму розподілу ознаки фізична активність в залежності від стадій діабету. Проаналізувавши дану діаграму можна зробити висновки:

- фізично активні пацієнти мають діабет у 3,5 раза менше, ніж інші;
- на показник перед діабету не впливає фізична активність.

Факт того, що фізично активні пацієнти мають діабет у 3,5 раза менше, ніж інші, свідчить про те, що фізична активність може впливати на ризик розвитку діабету. Зазвичай фізична активність сприяє поліпшенню чутливості до інсуліну та контролю рівня цукру в крові, а також допомагає у підтримці нормальної маси тіла.

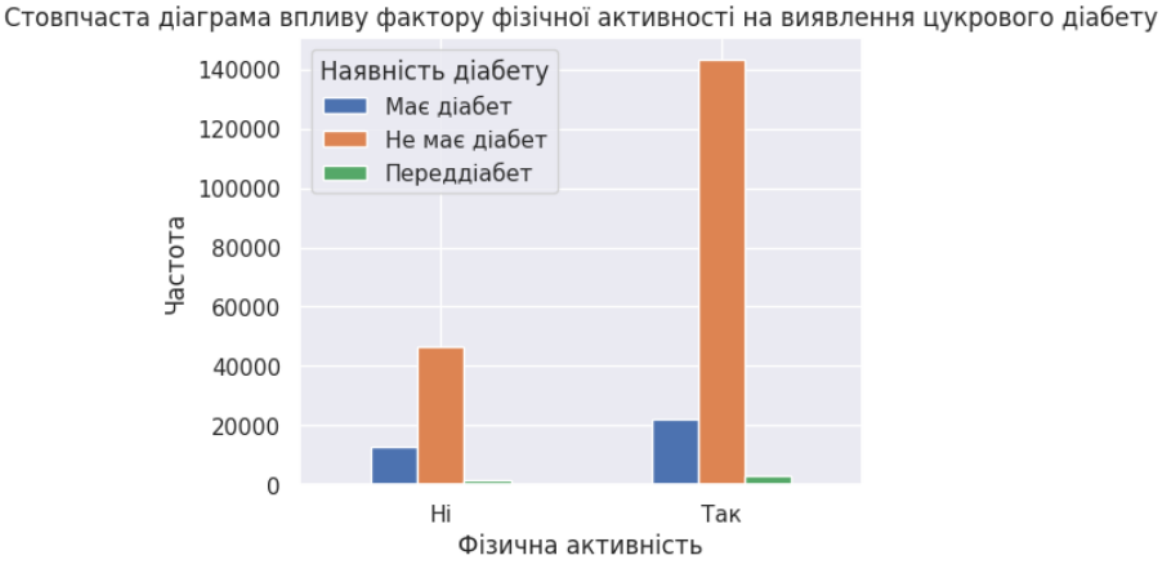


Рисунок 2.23 – Стовпчаста діаграма розподілу ознаки фізична активність в залежності від стадій діабету

Висновок:

Отже, можна зробити висновок, що є як дуже важливі фактори на виникнення діабету, а також – які можуть майже не вплинути на розвиток цукрового діабету. Тому проаналізувавши діаграму кореляції та діаграми розподіл ознак в залежності від стадій діабету (не має діабету, є діабет, перед діабет) визначили, що найбільший вплив мають оцінка здоров'я, артеріальний тиск, холестерин, індекс маси тіла, ішемічна хвороба серця, фізичне здоров'я, труднощі з ходьбою та вік.

Далі розглянули та проаналізували як впливає кожен цей показник на виявлення діабету та зробили наступні висновки:

– пацієнти, які мають ішемічну хворобу серця (ІХС), хворіють приблизно у 4 рази більше на цукровий діабет;

– лише 27% пацієнтів курять, а з них 8% мають цукровий діабет або перед діабет;

– жінки-пацієнти, які не мають діабету, на 20000 більше, ніж чоловіки.

Оскільки, жінки протягом життя мають різні гормональні зміни (вагітність, менопауза та інші);

– пацієнти, які не вживають фрукти, майже у 2 рази більше не хворіють на цукровий діабет. Свідчить про то, що не всі фрукти можуть нести корисні речовини та деякі з них містять багато цукру;

– овочі можуть бути важливим фактором для профілактики діабету, але важливо дотримуватися збалансованого споживання;

– пацієнти з високим холестерином мають у два рази частіше цукровий діабет. Оскільки, високий АТ може сприяти пошкодженню судин, погіршувати контроль рівня цукру в крові та підвищувати ризик серцево-судинних захворювань;

– пацієнти у віці від 50 до 64 років складають найбільшу кількість тих, хто має діабет. Оскільки, відбуваються нормальні процеси старіння, зниження функції підшлункової залози та збільшення інсулінорезистентності;

– кількість пацієнтів з ожиріння та цукровим діабетом складає 50%;

– частина пацієнтів, які мають інсульт та діабет, складає лише 1,5%;

– фізично активні пацієнти мають діабет у 3,5 раза менше, ніж інші. Це підкреслює важливість фізичної активності для зменшення ризику розвитку діабету.

Таким чином, ризик розвитку цукрового діабету визначається багатьма факторами, включаючи стать, вік, споживання фруктів та овочів, фізичну активність, рівень холестерину, наявність інших захворювань, таких як ішемічна хвороба серця та інсульт, а також генетичні особливості. Такі фактори як наявність медичного страхування, відвідування лікаря для профілактики, дохід, освіта не впливають на виявлення цукрового діабету взагалі.

2.7 Прогнозування виявлення цукрового діабету методами машинного навчання

У даному пункті було зроблено прогнозування виявлення цукрового діабету до його появи аби запобігти його розвитку в організмі людини за допомогою 4 методів машинного навчання: Decision Tree, Random Forest, K-NN, Ada Boost.

2.7.1 Прогнозування виявлення цукрового діабету методом K-NN

У даному пункті було зроблено прогнозування виявлення цукрового діабету методом K-NN у програмному середовищі Colab. Лістинг програми наведено у додатку Е.

Тоді перейдемо до результатів тренування та тестування наших моделей, де робота моделі була з наступними гіперпараметрами: кількість сусідів дорівнювала 5, а інші параметри за замовченням. Результати тестування моделі K-NN Classifier наведені в таблиці 2.1.

Таблиця 2.1

Результати моделі K-NN Classifier

class	precision	recall	f1-score	accuracy
0(не має діабет)	0.91	0.98	0.94	0.9
1(перед діабет)	0.00	0.00	0.00	
2 (має діабет)	0.31	0.10	0.15	

З таблиці 2.1 можна зробити наступні висновки:

- модель здійснила не збалансовану класифікацію даних, бо показники precision, recall, f1-score доволі різні для трьох класів, а це вказує на те, що модель не навчилася розрізняти ці три класи.

Тому потрібно було зробити балансування даних. Це було зроблено

за допомогою методу SMOTEENN, після цього дані стали збалансовані, де кожен клас мав таку кількість значення: : 0 – 27948 значення, 1 – 39597, 2 – 35628 значень.

Тоді ще раз було проведено тренування та тестування моделі, результати наведено у таблиці 2.2.

Таблиця 2.2

Результати моделі K-NN Classifier

class	precision	recall	f1-score	accuracy
0(не має діабет)	0.99	0.87	0.93	0.96
1(перед діабет)	0.96	1.00	0.98	
2 (має діабет)	0.94	0.98	0.96	

З таблиці 2.2 можна зробити наступні висновки:

– модель здійснила стабільну та збалансовану класифікацію даних, адже показники precision, f1-score доволі близькі для трьох класів, а це вказує на те, що модель добре розрізняє ці три класи;

– показник recall трохи відрізняються у нульовому класі це свідчить про модель для даного класу може пропускати позитивні приклади, які правильно були визначені, тобто має помилки.

Узагальнюючи, модель має високу точність для класів "перед діабет" і "має діабет", але меншу повноту для класу "не має діабет." Це означає, що модель може бути корисною для точної класифікації пацієнтів з діабетом і перед діабетом, але може не виявити всіх пацієнтів без діабету. Точність моделі в цілому є високою.

2.7.2 Прогнозування виявлення цукрового діабету методом Decision Tree

У даному пункті було зроблено прогнозування виявлення цукрового діабету методом Decision Tree у програмному середовищі Colab. Лістинг програми наведено у додатку Е.

Тоді перейдемо до результатів тренування та тестування наших моделей, де робота моделі була з наступними гіперпараметрами: criterion= 'entropy', max_depth=40, а інші параметри за замовченням. Результати тестування моделі Decision Tree Classifier наведені в таблиці 2.3.

Таблиця 2.3

Результати моделі Decision Tree Classifier

class	precision	recall	f1-score	accuracy
0(не має діабет)	0.92	0.89	0.91	0.92
1(перед діабет)	0.93	0.95	0.94	
2 (має діабет)	0.89	0.89	0.89	

З таблиці 2.3 можна зробити наступні висновки:

– для класу "Не має діабету" (клас 0) модель досягла високої точності та повноти (precision - 0.92, recall - 0.89), що вказує на високу здатність моделі правильно визначати відсутність діабету. Значення F1-середнього для цього класу становить 0.91, що є високим показником;

– для класу "Пред діабет" (клас 1) модель також показує високу точність та повноту (precision - 0.93, recall - 0.95). Значення F1-середнього для цього класу становить 0.94, що вказує на ефективність моделі в правильному розрізненні пред діабету;

– Для класу "Має діабет" (клас 2) модель виявляє більш низьку точність та повноту (precision - 0.89, recall - 0.89) порівняно з іншими класами. Значення F1-середнього для цього класу також становить 0.89.

Загальна точність моделі для всіх класів дорівнює 0.92, що свідчить про її загальну ефективність у класифікації. Модель має високі показники precision та recall для класів "Не має діабету" і "Пред діабет", що робить її корисною для виявлення цих станів у пацієнтів. Однак для класу "Має діабет" її ефективність менша, але все одно прийнятна.

2.7.3 Прогнозування виявлення цукрового діабету методом Random Forest

У даному пункті було зроблено прогнозування виявлення цукрового діабету методом Random Forest у програмному середовищі Colab. Лістинг програми наведено у додатку Е.

Тоді перейдемо до результатів тренування та тестування наших моделей, де робота моделі була з наступними гіперпараметрами: n_estimators=100, max_features=16, max_depth=16, а інші параметри за замовченням. Результати тестування моделі Random Forest Classifier наведені в таблиці 2.4.

Таблиця 2.4

Результати моделі Random Forest Classifier

class	precision	recall	f1-score	accuracy
0(не має діабет)	0.93	0.89	0.91	0.95
1(перед діабет)	0.86	0.86	0.86	
2 (має діабет)	0.79	0.82	0.81	

З таблиці 2.4 можна зробити наступні висновки:

– для класу "Не має діабету" (клас 0) модель демонструє високу точність (precision - 0.93) та достатню повноту (recall - 0.89). Значення F1-середнього для цього класу становить 0.91, що свідчить про добру здатність моделі правильно класифікувати відсутність діабету.

– для класу "Пред діабет" (клас 1) модель показує помірну точність (precision - 0.86) і повноту (recall - 0.86). Значення F1-середнього для цього класу становить 0.86, що вказує на середню ефективність моделі в класифікації пред діабету.

– для класу "Має діабет" (клас 2) модель показує найнижчі значення точності (precision - 0.79) і повноти (recall - 0.82) серед усіх класів. Значення F1-середнього для цього класу становить 0.81, що вказує на помірну ефективність моделі в правильному визначенні діабету.

Загальна точність моделі для всіх класів становить 0.95, що свідчить про її загальну ефективність у класифікації. Модель має найкращі показники для класу "Не має діабету", що робить її корисною для виявлення цього стану. Однак для класів "Пред діабет" і "Має діабет" точність і повнота менші, що може вказувати на більше помилкових класифікацій для цих груп.

2.7.4 Прогнозування виявлення цукрового діабету методом Ada Boost

У даному пункті було зроблено прогнозування виявлення цукрового діабету методом Ada Boost у програмному середовищі Colab. Лістинг програми наведено у додатку Е. Тоді перейдемо до результатів тренування та тестування наших моделей, де робота моделі була з наступними гіперпараметрами: n_estimators=100, max_features=16, max_depth=16, а інші параметри за замовченням. Результати тестування моделі Ada Boost Classifier наведені в таблиці 2.5.

Таблиця 2.5

Результати моделі Ada Boost Classifier

class	precision	recall	f1-score	accuracy
0(не має діабет)	0.84	0.77	0.80	0.63
1(перед діабет)	0.59	0.55	0.57	
2 (має діабет)	0.54	0.62	0.58	

З таблиці 2.5 можна зробити наступні висновки:

– для класу "Не має діабету" (клас 0) модель демонструє рівень точності (precision - 0.84) та повноти (recall - 0.77), що вказує на її здатність правильно класифікувати відсутність діабету. Значення F1-середнього для цього класу становить 0.80, що є високим результатом;

– для класу "Пред діабет" (клас 1) модель показує значення точності (precision - 0.59) та повноти (recall - 0.55), які не є дуже високими, вказуючи на помилкові класифікації пред діабету. Значення F1-середнього для цього класу становить 0.57, що є середнім результатом;

– для класу "Має діабет" (клас 2) модель показує значення точності (precision - 0.54) та повноти (recall - 0.62), які також не є дуже високими, вказуючи на помилкові класифікації діабету. Значення F1-середнього для цього класу становить 0.58, що є середнім результатом.

Загальна точність моделі для всіх класів становить 0.63, що свідчить про її загальну ефективність у класифікації. Проте точність та повнота для класів "Пред діабет" і "Має діабет" є недостатньо високими, вказуючи на помилкові класифікації для цих груп. Загальний рівень точності (accuracy) може бути покращений для покращення ефективності моделі.

2.8 Порівняння результатів прогнозування виявлення цукрового діабету методами машинного навчання

Після проведення прогнозування виявлення цукрового діабету в організмі людини, використовуючи показники та звички людей чотирма методами машинного навчання, можна зробити висновок, що найбільшу точність 95% має метод Random Forest, а найменшу 63% – Ada Boost, проте не треба одразу поспішати та обирати найкращою моделлю Random Forest Classifier. Оскільки, якщо порівняти показники precision, recall та f1-score, можна побачити, що

модель Decision Tree Classifier має найвищі показники f1-score для всіх трьох класів (0.94, 0.93, 0.89). Це вказує на кращу здатність моделі розрізняти всі три класи (не має діабету, перед діабет, має діабет) порівняно з іншими результати моделей. Тому найкращим методом для розв'язання поставленої задачі даної кваліфікаційної роботи є Decision Tree.

Висновки до розділу

У даному розділі було попереднє оброблено початкові дані, підібрані гіперпараметри для прогнозування методами машинного навчання (Decision Tree, Random Forest, K-Nearest Neighbors та Ada Boost), побудовано візуалізації розподілу ознак в залежності від стадій діабету (не має діабету, перед діабет та має діабет) та зроблено прогнозування й обрано найкращий результат моделі, яка була навчена одним з чотирьох запропонованих.

У рамках даної кваліфікаційної роботи було потрібно визначити причини, які найбільш всього впливають на розвиток діабету до його появи, а саме: АТ, рівень холестерин, фізична активність, індекс маси тіла, ішемічна хвороба серця, фізичне здоров'я, оцінка здоров'я та вік. Менш впливові – це споживання фруктів та овочів. Також визначилися, що куріння та психологічне здоров'я майже не впливає на розвиток діабету, а фактори які взагалі не доцільні – це освіта, дохід та наявність медичного страхування.

Далі було зроблено прогнозування та порівняно результати чотирьох моделей класифікації: Decision Tree, Random Forest, K-Nearest Neighbors та Ada Boost. Серед них найкращою моделлю виявилася – Decision Tree Classifier, який має точність 92%.

ВИСНОВКИ

Дана кваліфікаційна робота полягає у дослідженні аналізу причин та прогнозуванні виникнення діабету до його появи аби запобігти його розвитку методами машинного навчання. Дане дослідження є дуже важливим, оскільки, цукровий діабет є одна з невиліковних та серйозних захворювань, воно впливає на рівень цукру (глюкози) в крові та обумовлено генетичними, середовищними та стильовими факторами.

Для розв’язання даної проблеми використовувався набір даних, який був дуже не збалансованим, що додало певних складнощів. Тому для підготовки набору даних до навчання були здійснені такі кроки: видалення викидів за допомогою методу Isolated Forest, балансування класів методом SMOTEENN, розбиття даних на навчальну (70%) та тестову (30%) вибірки та масштабування атрибутів за допомогою методу Standard Scaler.

Наступним кроком було проведено аналіз причин виникнення цукрового діабету в організмі людини за допомогою візуалізації, де було визначено, що найбільш впливовими факторами є : АТ, рівень холестерин, фізична активність, індекс маси тіла, ішемічна хвороба серця, фізичне здоров’я, оцінка здоров’я та вік. Менш впливові – це споживання фруктів та овочів, бо деколи при неправильному споживанні чи у великій кількості призводить до зворотної дії впливу на організм. Також визначилися, що куріння та психологічне здоров’я майже не впливає на розвиток діабету.

Далі було зроблено прогнозування та порівняно результати чотирьох моделей класифікації: Decision Tree, Random Forest, K-Nearest Neighbors та Ada Boost. Серед них найкращою моделлю виявилася – Decision Tree Classifier, який має точність 92%.

Отримані результати мають велике значення, оскільки вони можуть бути використані для покращення роботи медичних фахівців у виявленні цукрового діабету до його появи аби запобігти його розвитку в організмі людини та підвищити шанси порятунку життя та здоров'я пацієнтів, які стикаються з цією серйозною та невиліковною хворобою.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Що таке штучний інтелект: Історія, види та складові. URL: https://gigacloud.ua/blog/navchannja/scho-take-shtuchnij-intelekt-istorija-vidi-ta-skladovi (дата звернення: 01.07.2023 року)

2. Від Ш до І: що таке штучний інтелект та як він трансформує світ. URL: https://speka.media/ai/vid-s-do-i-shho-take-stucnii-intelekt-ta-yak-vin-transformuje-svit-xv7039 (дата звернення: 10.07.2023 року)

3. Як діє штучний інтелект і перспективи його використання.. URL: https://aiconference.com.ua/uk/news/printsipi-raboti-iskusstvennogo-intellekta-i-perspektiva-ego-ispolzovaniya-92238 (дата звернення: 11.07.2023 року)

4. Поява та перспективи розвитку штучного інтелекту. URL: https://duikt.edu.ua/ua/news-1-576-8835-poyava-ta-perspektivi-rozvitku-shtuchnogo-intelektu_kafedra-shtuchnogo-intelektu (дата звернення: 16.08.2023 року)

5. Штучний інтелект, машинне навчання та нейронні мережі: в чому різниця і для чого їх використовують. URL: https://evergreens.com.ua/ua/articles/machine-learning-overview.html (дата звернення: 01.09.2023 року)

6. Що таке machine learning ? Як працює машинне навчання та де воно використовується. URL: https://www.telegraf.in.ua/advertisement/10119869-scho-take-machine-learning-jak-pracjuye-mashinne-navchannja-ta-de-vono-vikoristovuyetsja.html (дата звернення: 25.06.2023 року)

7. Машинне навчання простими словами. Частина 1. URL: http://www.mmf.lnu.edu.ua/ar/1739 (дата звернення: 17.07.2023 року)

8. Оцінка якості моделі класифікації. URL: https://studfile.net/preview/9974842/page:22/ (дата звернення: 22.08.2023 року)

9. Оцінка якості роботи алгоритму машинного навчання. URL: https://stud.com.ua/139975/informatika/otsinka_yakosti_roboti_algoritmu_mashinnogo_navchannya (дата звернення: 07.07.2023 року)

10. Дерево рішень. URL: https://ua5.org/algorithm/1976-derevo-rishen.html (дата звернення: 13.08.2023 року)

11. Decision Tree. URL: https://www.geeksforgeeks.org/decision-tree/_(дата звернення: 19.08.2023 року)

12. Decision Trees. URL: https://scikit-learn.org/stable/modules/tree.html (дата звернення: 01.07.2023 року)

13. Дерева прийняття рішень. URL: https://stud.com.ua/139987/informatika/dereva_priynyattya_rishen_(дата звернення: 12.07.2023 року)

14. Випадковий ліс (Random Forest) URL: https://alexanderdyakonov.wordpress.com/2016/11/14/%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D1%8B%D0%B9-%D0%BB%D0%B5%D1%81-random-forest/ (дата звернення: 14.08.2023 року)

15. Random Forest Algorithm. URL: https://www.javatpoint.com/machine-learning-random-forest-algorithm (дата звернення: 07.08.2023 року)

16. K-Nearest Neighbor (KNN) Algorithm for Machine Learning. URL: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (дата звернення: 01.07.2023 року)

17. Метод «найближчого сусіда» або системи міркувань на основі аналогічних випадків. URL: https://studfile.net/preview/7818687/page:2/(дата звернення: 11.08.2023 року)

18. AdaBoost Algorithm. URL: analyticsvidhya.com (дата звернення: 17.08.2023 року)

19. Improving Machine Learning Diabetes Prediction Models for the Utmost Clinical Effectiveness. URL: https://www.mdpi.com/2075-4426/12/11/1899 (дата звернення: 22.06.2023 року)

20. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. URL: https://www.sciencedirect.com/science/article/pii/S0169260722001596 (дата звернення: 17.06.2023 року)

21. Посилання

22. Кількість діабетиків у світі до 2050 року може зрости майже втричі. URL: https://thepage.ua/ua/news/kilkist-diabetikiv-v-sviti-mozhe-zrosti-do-13-milyarda-lyudej-do-2050-roku (дата звернення: 19.08.2023 року)

23. RSV Prevention. URL: https://www.cdc.gov/ (дата звернення: 25.07.2023 року)

24. Outliers in Machine Learning. URL: https://medium.com/analytics-vidhya/outliers-in-machine-learning-e830b2bd8660 (дата звернення: 19.08.2023 року)

25. Isolated Forest. URL: https://medium.com/@corymaklin/isolation-forest-799fceacdda4 (дата звернення: 19.08.2023 року)

26. Категорійна змінна. URL: https://uk.wikipedia.org/wiki/%D0%9A%D0%B0%D1%82%D0%B5%D0%B3%D0%BE%D1%80%D1%96%D0%B9%D0%BD%D0%B0_%D0%B7%D0%BC%D1%96%D0%BD%D0%BD%D0%B0 (дата звернення: 06.07.2023 року)

27. Imbalanced Classification in Python^ SMOTE-ENN method. URL: https://towardsdatascience.com/imbalanced-classification-in-python-smote-enn-method-db5db06b8d50 (дата звернення: 10.08.2023 року)

28. Standard Scaler. URL: learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (дата звернення: 19.07.2023 року)

29. Parameters and Hyperperemeters in Machine Learning and Deep Learning. URL: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac (дата звернення: 14.08.2023 року)

30. Hyperperemeters. URL: https://www.techopedia.com/definition/34625/hyperparameter-ml-hyperparameter (дата звернення: 03.07.2023 року)

31. Decision Tree Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (дата звернення: 05.07.2023 року)

32. Random Forest Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (дата звернення: 11.07.2023 року)

33. K-Neighbors Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (дата звернення: 09.07.2023 року)

34. Ada Boost Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html (дата звернення: 13.08.2023 року)

Додаток А. Відомість матеріалів комплексної кваліфікаційної роботи

№ з/п	Позначення	Назва	Кількість	Примітки
1				
2		Документація		
3				
4	САУ.КР.22.16.ПЗ	Пояснювальна записка	81	Формат А4
5				
6		Демонстраційні матеріали	1	Презентація на CD-R
7				
8		Копія роботи	1	Диск CD-R
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				

					САіУ.КР.22.16.ДА.ПЗ.			
Змін.	Аркуш	№ докум.	Підпис	Дата				
Розроб.		Сидоренко К.В.			**Матеріали кваліфікаційної роботи**	Літ.	Аркуш	Аркушів
К. розд.		Хом'як Т.В.						
Керівн.		Хом'як Т.В.				НТУ «ДП» 12; 124м-22-1		
Н.контр.		Хом'як Т.В.						
Зав. каф.		Желдак Т.А.						

Додаток Б. Відгук на кваліфікаційну роботу бакалавра

Відгук

на кваліфікаційну роботу магістра

студентки групи *124м-22-1 Сидоренко Катерини Віталіївни*

спеціальності *124 Системний аналіз*

Тема комплексної кваліфікаційної роботи: Аналіз причин та прогнозування виявлення цукрового діабету методами машинного навчання.

Обсяг комплексної кваліфікаційної роботи 81 с., 30 рисунків, 6 таблиць, 7 додатків, 34 джерела.

Мета кваліфікаційної роботи: аналіз причин виникнення діабету за допомогою візуалізації та прогнозування виникнення цукрового діабету методами машинного навчання для того, щоб зменшити кількість випадків виникнення діабету.

Актуальність дослідження полягає у тому, що цукровий діабет є серйозне та невиліковне захворювання, яке може призвести до численних ускладнень та проблем зі здоров'ям. Тому потрібно розробити модель для раннього та більш правильного прогнозування виявлення діабету.

Тема кваліфікаційної роботи безпосередньо пов'язана з об'єктом діяльності магістра спеціальності «Системний аналіз», оскільки включає в себе розробку програмного середовища на мові програмування Python для аналізу за допомогою візуалізації та створенні прогнозу методами машинного навчання.

Виконані в кваліфікаційній роботі завдання відповідають вимогам до професійної діяльності фахівця освітньо-кваліфікаційного рівня магістра. Оригінальність наукових рішень полягає у використанні методів машинного навчання: Decision Tree, Random Forest, K-NN, Ada Boost для прогнозування з використанням оцінки якості моделей класифікації.

Практичне значення результатів кваліфікаційної роботи полягає у виборі найкращої моделі класифікації для раннього (до його появи) діагностування цукрового діабету

Висновки підтверджують можливість використання результатів роботи у подальших наукових дослідження або ж використання у медичних центрах для виявлення цукрового діабету до його появи аби запобігти його розвитку.

Оформлення пояснювальної записки та демонстраційного матеріалу до неї виконано згідно з вимогами. Роботу виконано самостійно, відповідно до завдання та у повному обсязі.

Кваліфікаційна робота в цілому заслуговує на оцінки: відмінно.

З урахуванням висловлених зауважень автор заслуговує присвоєння кваліфікації «магістр з системного аналізу».

Керівник кваліфікаційної роботи бакалавра,

К.ф.-м.н., доцент _____ Хом'як Т.В.

Додаток В. Рецензія на кваліфікаційну роботу магістра

Рецензія

На кваліфікаційну роботу магістра

студентки групи *124м-22-1 Сидоренко Катерини Віталіївни*

спеціальності *124 Системний аналіз*

Тема комплексної кваліфікаційної роботи: Аналіз причин та прогнозування виявлення цукрового діабету методами машинного навчання.

Обсяг комплексної кваліфікаційної роботи 81 с., 30 рисунків, 6 таблиць, 7 додатків, 34 джерела.

Висновок про відповідність кваліфікаційної роботи завданню та освітньо-професійній програмі спеціальності – кваліфікаційна робота відповідає вимогам до професійної діяльності фахівця освітньо-кваліфікаційного рівня магістра спеціальності 124 Системний аналіз і управління.

Зміст пояснювальної записки відповідає темі кваліфікаційної роботи.

Загальна характеристика кваліфікаційної роботи, ступінь використання нормативно-методичної літератури та передового досвіду. Кваліфікаційна робота містить 2 розділи. В інформаційно-аналітичному розділі проведено огляд методів машинного навчання їх особливості та складові й оцінки якості моделей класифікації. У спеціальному розділі було проаналізовано причини виникнення цукрового діабету, зроблено прогноз чотирма методами машинного навчання для визначення найкращої моделі класифікації для того, щоб можна було діагностувати цукровий діабет до його появи.

Позитивні сторони кваліфікаційної роботи: детальний аналіз причин виникнення діабету за допомогою візуалізації на мові програмування Python.

Основні недоліки кваліфікаційні роботи: немає.

Рекомендації: дослідити та зробити прогнозування всіма можливими методами машинного навчання для того, щоб обрати найкращий метод, який можна вводити до системи медичних закладів.

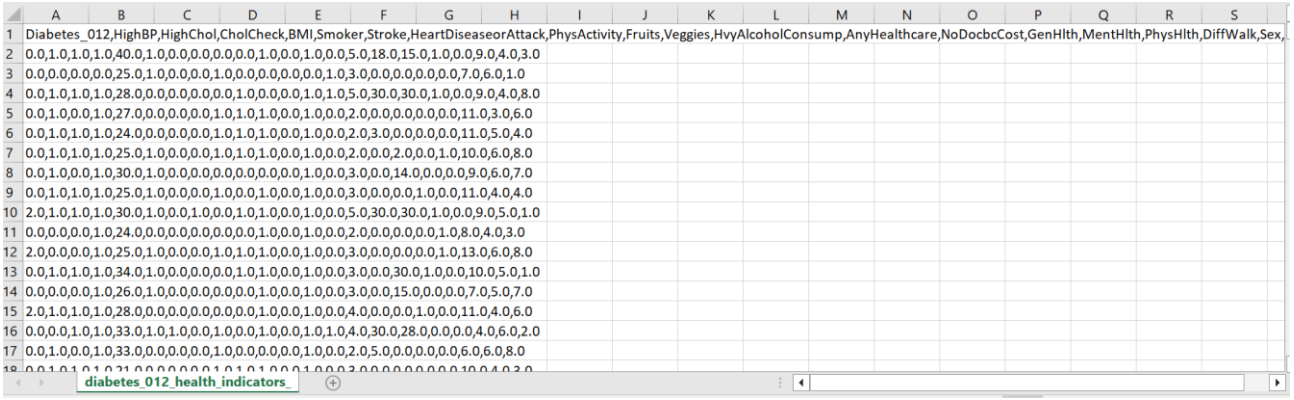
Кваліфікаційна робота в цілому заслуговує на оцінки: відмінно.

З урахуванням висловлених зауважень автор заслуговує присвоєння кваліфікації «магістр з системного аналізу».

Рецензент,

К.ф.-м _____ ПІБ.

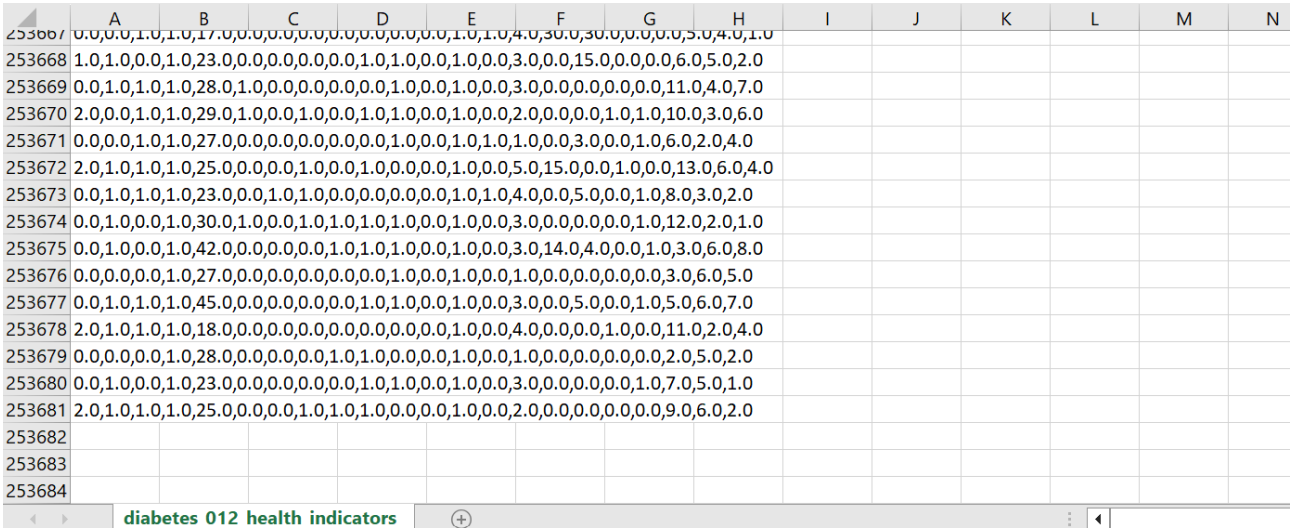
Додаток Г. Початкові дані



	A
1	Diabetes_012,HighBP,HighChol,CholCheck,BMI,Smoker,Stroke,HeartDiseaseorAttack,PhysActivity,Fruits,Veggies,HvyAlcoholConsump,AnyHealthcare,NoDocbcCost,GenHlth,MentHlth,PhysHlth,DiffWalk,Sex,
2	0.0,1.0,1.0,1.0,40.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,5.0,18.0,15.0,1.0,0.0,9.0,4.0,3.0
3	0.0,0.0,0.0,0.0,25.0,1.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,1.0,3.0,0.0,0.0,0.0,0.0,7.0,6.0,1.0
4	0.0,1.0,1.0,1.0,28.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,1.0,5.0,30.0,30.0,1.0,0.0,9.0,4.0,8.0
5	0.0,1.0,0.0,1.0,27.0,0.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,2.0,0.0,0.0,0.0,0.0,11.0,3.0,6.0
6	0.0,1.0,1.0,1.0,24.0,0.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,2.0,3.0,0.0,0.0,0.0,11.0,5.0,4.0
7	0.0,1.0,1.0,1.0,25.0,1.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,2.0,0.0,2.0,0.0,1.0,10.0,6.0,8.0
8	0.0,1.0,0.0,1.0,30.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,3.0,0.0,14.0,0.0,0.0,9.0,6.0,7.0
9	0.0,1.0,1.0,1.0,25.0,1.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,1.0,0.0,3.0,0.0,0.0,0.0,1.0,0.0,11.0,4.0,4.0
10	2.0,1.0,1.0,1.0,30.0,1.0,0.0,0.1,0.0,0.1,1.0,0.0,0.1,0.0,0.5,0.30.0,30.0,1.0,0.0,9.0,5.0,1.0
11	0.0,0.0,0.0,1.0,24.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,2.0,0.0,0.0,0.0,0.0,1.0,8.0,4.0,3.0
12	2.0,0.0,0.0,1.0,25.0,1.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,3.0,0.0,0.0,0.0,0.0,1.0,13.0,6.0,8.0
13	0.0,1.0,1.0,1.0,34.0,1.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,1.0,0.0,3.0,0.0,30.0,1.0,0.0,0.10.0,5.0,1.0
14	0.0,0.0,0.0,1.0,26.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,3.0,0.0,15.0,0.0,0.0,0.7.0,5.0,7.0
15	2.0,1.0,1.0,1.0,28.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,4.0,0.0,0.0,1.0,0.0,11.0,4.0,6.0
16	0.0,0.0,1.0,1.0,33.0,1.0,1.0,0.0,0.1,0.0,0.0,1.0,0.0,1.0,1.0,4.0,30.0,28.0,0.0,0.0,4.0,6.0,2.0
17	0.0,1.0,0.0,1.0,33.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,1.0,0.0,2.0,5.0,0.0,0.0,0.0,6.0,6.0,8.0
18	0.0,1.0,1.0,1.0,21.0,0.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,3.0,0.0,0.0,0.0,0.0,10.0,4.0,3.0

diabetes_012_health_indicators_

Рисунок Г.1 – Початкові дані для аналізу та прогнозування виявлення цукрового діабету до його появи аби запобігти його розвитку в організмі людини



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
253667	0.0,0.0,1.0,1.0,17.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,4.0,30.0,30.0,0.0,0.0,5.0,4.0,1.0													
253668	1.0,1.0,0.0,1.0,23.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,1.0,0.0,3.0,0.0,15.0,0.0,0.0,6.0,5.0,2.0													
253669	0.0,1.0,1.0,1.0,28.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,3.0,0.0,0.0,0.0,0.0,11.0,4.0,7.0													
253670	2.0,0.0,1.0,1.0,29.0,1.0,0.0,1.0,0.0,1.0,1.0,0.0,1.0,0.0,2.0,0.0,0.0,1.0,1.0,10.0,3.0,6.0													
253671	0.0,0.0,1.0,1.0,27.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,1.0,1.0,0.0,3.0,0.0,1.0,6.0,2.0,4.0													
253672	2.0,1.0,1.0,1.0,25.0,0.0,0.0,1.0,0.0,1.0,0.0,0.0,1.0,0.0,5.0,15.0,0.0,1.0,0.0,13.0,6.0,4.0													
253673	0.0,1.0,1.0,1.0,23.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,4.0,0.0,5.0,0.0,1.0,8.0,3.0,2.0													
253674	0.0,1.0,0.0,1.0,30.0,1.0,0.0,1.0,1.0,1.0,1.0,0.0,1.0,0.0,3.0,0.0,0.0,0.0,1.0,12.0,2.0,1.0													
253675	0.0,1.0,0.0,1.0,42.0,0.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,3.0,14.0,4.0,0.0,1.0,3.0,6.0,8.0													
253676	0.0,0.0,0.0,1.0,27.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,3.0,6.0,5.0													
253677	0.0,1.0,1.0,1.0,45.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,3.0,0.0,5.0,0.0,1.0,5.0,6.0,7.0													
253678	2.0,1.0,1.0,1.0,18.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,4.0,0.0,0.0,1.0,0.0,11.0,2.0,4.0													
253679	0.0,0.0,0.0,1.0,28.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0,2.0,5.0,2.0													
253680	0.0,1.0,0.0,1.0,23.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,1.0,0.0,3.0,0.0,0.0,0.0,1.0,7.0,5.0,1.0													
253681	2.0,1.0,1.0,1.0,25.0,0.0,0.0,1.0,1.0,1.0,0.0,0.0,1.0,0.0,2.0,0.0,0.0,0.0,0.0,9.0,6.0,2.0													
253682														
253683														
253684														

diabetes 012 health indicators

Рисунок Г.2 – Початкові дані для аналізу та прогнозування виявлення цукрового діабету до його появи аби запобігти його розвитку в організмі людини

Додаток Д. Матриця кореляції причин впливу на виникнення цукрового діабету

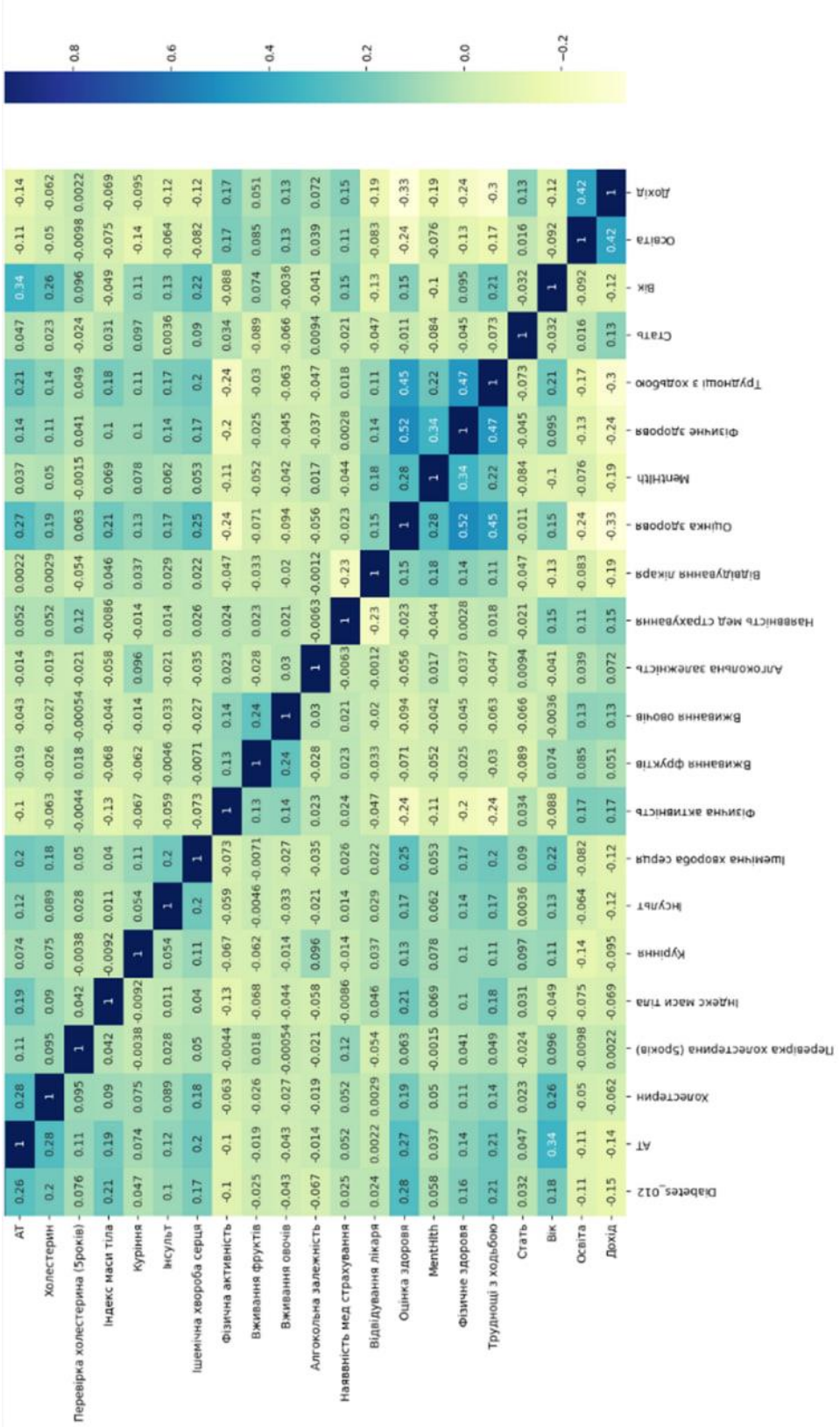


Рисунок Д.1 – Матриця кореляції для аналізу виявлення цукрового діабету до його появи аби запобігти його розвитку в організмі людини

Додаток Е Лістинг програмного продукту для аналізу та прогнозування виявлення цукрового діабету

```
pip install sklearn

pip install scikit-learn

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

from sklearn import linear_model

from sklearn.linear_model import LogisticRegression
from sklearn.tree  import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier

from sklearn import tree
from sklearn.metrics import accuracy_score
#from scikitplot.metrics import plot_roc_curve
#from sklearn.metrics import plot_roc_curve
from sklearn.metrics import f1_score,  recall_score, precision_score
from sklearn.metrics import RocCurveDisplay #замість строки вище
from sklearn.model_selection import cross_val_predict
from sklearn.model_selection import cross_val_score

from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split

from sklearn.feature_selection import SelectPercentile
from sklearn.feature_selection import chi2 , f_classif

from sklearn.metrics import mean_absolute_error,r2_score,mean_squared_error
from sklearn.feature_selection import RFE
warnings.filterwarnings('ignore')

abosalah="diabetes_012_health_indicators_BRFSS2015.csv"
data= pd.read_csv(abosalah)
data.head()
data

"""
Diabetes_012
0 = no diabetes 1 = prediabetes 2 = diabetes
Diabetes_012
0 = відсутність діабету 1 = перед діабет 2 = діабет
HighBP
```

0 = no high BP 1 = high BP

HighBP

0 = немає високого АТ 1 = високий АТ

HighChol

0 = no high cholesterol 1 = high cholesterol

HighChol

0 = відсутність високого холестерину 1 = високий рівень холестерину

CholCheck

0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 yearsCholCheck

0 = відсутність холестерину перевірка через 5 років 1 = так перевірка холестерину через 5 років

BMI

Body Mass Index

IMT

Індекс маси тіла

Smoker

Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes

Курець

Ви викурили хоча б 100 сигарет за все своє життя? [Примітка: 5 пачок = 100 сигарет] 0 = ні 1 = так

Stroke

(Ever told) you had a stroke. 0 = no 1 = yes

Обведення

(Коли-небудь говорив) у вас був інсульт. 0 = ні 1 = так

HeartDiseaseorAttack

coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yeseartDiseaseorAttack

ішемічна хвороба серця (ІХС) або інфаркт міокарда (ІМ) 0 = ні 1 = так

PhysActivity

physical activity in past 30 days - not including job 0 = no 1 = yes

Фізична активність

фізична активність за останні 30 днів - не включаючи роботу 0 = ні 1 = так

Fruits

Consume Fruit 1 or more times per day 0 = no 1 = yes

Фрукти

Споживайте фрукти 1 або більше разів на день 0 = ні 1 = так

Veggies

Consume Vegetables 1 or more times per day 0 = no 1 = yes

Овочі

Споживайте овочі 1 або більше разів на день 0 = ні 1 = так

HvyAlcoholConsump

Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes

HvyAlcoholConsump

Люди, які сильно п'ють (дорослі чоловіки, які вживають більше 14 напоїв на тиждень, і дорослі жінки, які вживають більше 7 напоїв на тиждень) 0 = ні 1 = так

AnyHealthcare

Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes

AnyHealthcare

Майте будь-яке медичне страхування, включаючи медичне страхування, передплачені плани, такі як HMO тощо. 0 = ні 1 = так

NoDocbcCost

Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes

NoDocbcВартість

Чи був час за останні 12 місяців, коли вам потрібно було звернутися до лікаря, але ви не могли через вартість? 0 = ні 1 = так

GenHlth

Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
GenHlth
Ви б сказали, що в цілому ваше здоров'я таке: шкала 1-5 1 = відмінна 2 = дуже хороша 3 = хороша 4 = справедлива 5 = погана
MentHlth
Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
MentHlth
Тепер думайте про своє психічне здоров'я, яке включає стрес, депресію та проблеми з емоціями, протягом скількох днів протягом останніх 30 днів ваше психічне здоров'я було поганим? масштаб 1-30 днів
PhysHlth
Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
PhysHlth
Тепер подумайте про своє фізичне здоров'я, яке включає фізичні захворювання та травми, протягом скількох днів протягом останніх 30 днів ваше фізичне здоров'я було поганим? масштаб 1-30 днів
DiffWalk
У вас є серйозні труднощі при ходьбі або підйомі по сходах? 0 = ні 1 = так
Sex
0 = female 1 = male
Age
13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or olderEducation

Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
Income
Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more
"""

```
data['Diabetes_012'].value_counts()

data.info()

data["Diabetes_012"] = data["Diabetes_012"].astype(int)
data["HighBP"] = data["HighBP"].astype(int)
data["HighChol"] = data["HighChol"].astype(int)
data["CholCheck"] = data["CholCheck"].astype(int)
data["BMI"] = data["BMI"].astype(int)
data["Smoker"] = data["Smoker"].astype(int)
data["Stroke"] = data["Stroke"].astype(int)
data["HeartDiseaseorAttack"] = data["HeartDiseaseorAttack"].astype(int)
data["PhysActivity"] = data["PhysActivity"].astype(int)
data["Fruits"] = data["Fruits"].astype(int)
data["Veggies"] = data["Veggies"].astype(int)
data["HvyAlcoholConsump"] = data["HvyAlcoholConsump"].astype(int)
data["AnyHealthcare"] = data["AnyHealthcare"].astype(int)
data["NoDocbcCost"] = data["NoDocbcCost"].astype(int)
data["GenHlth"] = data["GenHlth"].astype(int)
data["MentHlth"] = data["MentHlth"].astype(int)
data["PhysHlth"] = data["PhysHlth"].astype(int)
data["DiffWalk"] = data["DiffWalk"].astype(int)
data["Sex"] = data["Sex"].astype(int)
```

```
data["Age"] = data["Age"].astype(int)
data["Education"] = data["Education"].astype(int)
data["Income"] =data["Income"].astype(int)

data.info()

# df.shape
print(f"Num rows: {len(data)}")
print(f"Num columns: {len(data.columns)}")

data.shape

data['Diabetes_012'].value_counts(normalize=True).plot(kind='bar');

data.columns

data.describe()

data.isnull().sum()

data.isnull().sum().any()

data.duplicated().sum()

data.loc[data.duplicated(),:]

data.drop_duplicates(inplace=True)

data.shape

#data.rename(columns={'HighBP': 'AT', 'HighChol': 'Холестерин'}, inplace=True)

#data.rename(columns={'CholCheck': 'Перевірка холестерина (5років)', 'BMI': 'Індекс маси тіла', 'Smoker': 'Куріння', 'Stroke': 'Інсульт', 'HeartDiseaseorAttack': 'Ішемічна хвороба серця', 'PhysActivity': 'Фізична активність'}, inplace=True)
#data.rename(columns={'Fruits': 'Вживання фруктів', 'Veggies': 'Вживання овочів', 'HvyAlcoholConsump': 'Алгокольна залежність', 'AnyHealthcare': 'Наяввність мед страхування', 'NoDocbcCost': 'Відвідування лікаря', 'GenHlth': 'Оцінка здоровя'}, inplace=True)
#data.rename(columns={'MenthHlth': 'Психічне здоровя', 'PhysHlth': 'Фізичне здоровя', 'DiffWalk': 'Труднощі з ходьбою', 'Sex': 'Стать', 'Age': 'Вік', 'Education': 'Освіта','Income': 'Дохід'}, inplace=True)

plt.figure(figsize=(20,10))
sns.heatmap(data.corr(), annot=True, cmap="YlGnBu")

df_vis=data.copy()

#transform data
df_vis.Diabetes_012[df_vis['Diabetes_012'] == 0] = 'No Diabetes'
df_vis.Diabetes_012[df_vis['Diabetes_012'] == 1] = 'Pre Diabetes'
df_vis.Diabetes_012[df_vis['Diabetes_012'] == 2] = 'Diabetes'

df_vis.HighBP[df_vis['HighBP'] == 0] = 'No High'
df_vis.HighBP[df_vis['HighBP'] == 1] = 'High BP'
```

```
df_vis.HighChol[df_vis['HighChol'] == 0] = 'No High Cholesterol'
df_vis.HighChol[df_vis['HighChol'] == 1] = 'High Cholesterol'

df_vis.CholCheck[df_vis['CholCheck'] == 0] = 'No Cholesterol Check in 5 Years'
df_vis.CholCheck[df_vis['CholCheck'] == 1] = 'Cholesterol Check in 5 Years'

df_vis.Smoker[df_vis['Smoker'] == 0] = 'No'
df_vis.Smoker[df_vis['Smoker'] == 1] = 'Yes'

df_vis.Stroke[df_vis['Stroke'] == 0] = 'No'
df_vis.Stroke[df_vis['Stroke'] == 1] = 'Yes'

df_vis.HeartDiseaseorAttack[df_vis['HeartDiseaseorAttack'] == 0] = 'No'
df_vis.HeartDiseaseorAttack[df_vis['HeartDiseaseorAttack'] == 1] = 'Yes'

df_vis.PhysActivity[df_vis['PhysActivity'] == 0] = 'No'
df_vis.PhysActivity[df_vis['PhysActivity'] == 1] = 'Yes'

df_vis.Fruits[df_vis['Fruits'] == 0] = 'No'
df_vis.Fruits[df_vis['Fruits'] == 1] = 'Yes'

df_vis.Veggies[df_vis['Veggies'] == 0] = 'No'
df_vis.Veggies[df_vis['Veggies'] == 1] = 'Yes'

df_vis.HvyAlcoholConsump[df_vis['HvyAlcoholConsump'] == 0] = 'No'
df_vis.HvyAlcoholConsump[df_vis['HvyAlcoholConsump'] == 1] = 'Yes'

df_vis.AnyHealthcare[df_vis['AnyHealthcare'] == 0] = 'No'
df_vis.AnyHealthcare[df_vis['AnyHealthcare'] == 1] = 'Yes'

df_vis.NoDocbcCost[df_vis['NoDocbcCost'] == 0] = 'No'
df_vis.NoDocbcCost[df_vis['NoDocbcCost'] == 1] = 'Yes'
df_vis.GenHlth[df_vis['GenHlth'] == 1] = 'Excellent'
df_vis.GenHlth[df_vis['GenHlth'] == 2] = 'Very Good'
df_vis.GenHlth[df_vis['GenHlth'] == 3] = 'Good'
df_vis.GenHlth[df_vis['GenHlth'] == 4] = 'Fair'
df_vis.GenHlth[df_vis['GenHlth'] == 5] = 'Poor'

df_vis.DiffWalk[df_vis['DiffWalk'] == 0] = 'No'
df_vis.DiffWalk[df_vis['DiffWalk'] == 1] = 'Yes'

df_vis.Sex[df_vis['Sex'] == 0] = 'Female'
df_vis.Sex[df_vis['Sex'] == 1] = 'Male'

df_vis.Education[df_vis['Education'] == 1] = 'Never Attended School'
df_vis.Education[df_vis['Education'] == 2] = 'Elementary'
df_vis.Education[df_vis['Education'] == 3] = 'Some high school'
df_vis.Education[df_vis['Education'] == 4] = 'High school graduate'
df_vis.Education[df_vis['Education'] == 5] = 'Some college or technical school'
df_vis.Education[df_vis['Education'] == 6] = 'College graduate'

df_vis.Income[df_vis['Income'] == 1] = 'Less Than $10,000'
```

```
df_vis.Income[df_vis['Income'] == 2] = 'Less Than $10,000'
df_vis.Income[df_vis['Income'] == 3] = 'Less Than $10,000'
df_vis.Income[df_vis['Income'] == 4] = 'Less Than $10,000'
df_vis.Income[df_vis['Income'] == 5] = 'Less Than $35,000'
df_vis.Income[df_vis['Income'] == 6] = 'Less Than $35,000'
df_vis.Income[df_vis['Income'] == 7] = 'Less Than $35,000'
df_vis.Income[df_vis['Income'] == 8] = '$75,000 or More'

unique_values = {}
for col in df_vis.columns:
    unique_values[col] = df_vis[col].value_counts().shape[0]

pd.DataFrame(unique_values, index=['unique value count']).transpose()

cols = list(df_vis.columns)
cols_df=cols[1:]

plt.figure(figsize=(15,40))
for i in range(len(cols_df)):
    plt.subplot(8,3,i+1)
    plt.title(cols_df[i])
    plt.xticks(rotation=90)
    plt.hist(df_vis[cols_df[i]])

plt.tight_layout()

df_vis['Diabetes_012'].value_counts()

# pie plot of diabetes ratio співвідношення трьох можливих стадій діабету

plt.figure(figsize=(8,6))
labels = ['Не має діабет ', 'Є діабет', 'Перед діабет']
sizes = [df_vis['Diabetes_012'].value_counts()[0], df_vis['Diabetes_012'].value_counts()[1], df_vis['Diabetes_012'].value_counts()[2]]
colors = ['lightskyblue', 'grey', 'lightcoral']
explode = (0.05, 0.05, 0)  # explode 1st slice
plt.pie(sizes, explode=explode, labels=labels, autopct='%.1f%%', colors=colors, data = df_vis);

#data.rename(columns={'HighBP': 'AT', 'HighChol': 'Холестерин'}, inplace=True)

#data.rename(columns={'CholCheck': 'Перевірка холестерина (5років)', 'BMI': 'Індекс маси тіла', 'Smoker': 'Куріння', 'Stroke': 'Інсульт',
'HeartDiseaseorAttack': 'Ішемічна хвороба серця', 'PhysActivity': 'Фізична активність'}, inplace=True)
#data.rename(columns={'Fruits': 'Вживання фруктів', 'Veggies': 'Вживання овочів', 'HvyAlcoholConsump': 'Алкогольна залежність',
'AnyHealthcare': 'Наяввність мед страхування', 'NoDocbcCost': 'Відвідування лікаря', 'GenHlth': 'Оцінка здоровя'}, inplace=True)
#data.rename(columns={'MentHlth': 'Психічне здоровя', 'PhysHlth': 'Фізичне здоровя', 'DiffWalk': 'Труднощі з ходьбою', 'Sex': 'Стать', 'Age': 'Вік',
'Education': 'Освіта','Income': 'Дохід'}, inplace=True)



data.drop('Diabetes_012', axis=1).corrwith(data.Diabetes_012).plot(kind='bar', grid=True, figsize=(15, 6)
, title="Стовпчаста діаграма колеріції факторів, які впливають на виявлення цукрового діабету ",color="blue");

plt.figure(figsize=(12,4))
```

```
x= sns.countplot(x='Diabetes_012',data=df_vis,hue='Sex')
#plt.xticks(rotation=90)
plt.xlabel('Наявність діабету')
plt.ylabel('Кількість ')
plt.legend(title='Стать', labels=['Жінка', 'Чоловік'])
x.set_xticklabels(['Не має діабету', 'Має діабет', 'Перед діабет'])
plt.title('Стовпчаста діаграма розподілу діабету за статтю',fontdict={'fontsize':20})
for i in x.patches:
   x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()

#smoke
plt.figure(figsize=(12,5))
x= sns.countplot(x='Smoker', hue='Diabetes_012' , data = df_vis);
plt.xlabel('Куріння')
plt.ylabel('Кількість ')
plt.legend(title='Наявність діабету', labels=['Не має діабету', 'Має діабет', 'Перед діабет'])


x.set_xticklabels(['Так', 'Ні'])
plt.title('Стовпчаста діаграма впливу фактора куріння на виявлення цукрового діабету ',fontdict={'fontsize':20})
for i in x.patches:
   x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()

#smoke
plt.figure(figsize=(12,5))
x= sns.countplot(x='Stroke', hue='Diabetes_012' , data = df_vis);
plt.xlabel('Інсульт')
plt.ylabel('Кількість ')
plt.legend(title='Наявність діабету', labels=['Не має діабету', 'Має діабет', 'Перед діабет'])


x.set_xticklabels(['Ні', 'Так'])
plt.title('Стовпчаста діаграма впливу фактора інсульту на виявлення цукрового діабету ',fontdict={'fontsize':20})
for i in x.patches:
   x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()

#ішемічна хвороба серця
plt.figure(figsize=(12,5))
x= sns.countplot(x='HeartDiseaseorAttack', hue='Diabetes_012' , data = df_vis);
plt.xlabel('Ішемічна хвороба серця')
plt.ylabel('Кількість ')
plt.legend(title='Наявність діабету', labels=['Не має діабету', 'Має діабет', 'Перед діабет'])


x.set_xticklabels(['Так', 'Ні'])
plt.title('Стовпчаста діаграма впливу фактору ІХС на виявлення цукрового діабету ',fontdict={'fontsize':20})
for i in x.patches:
   x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()
```

```
#фрукти
plt.figure(figsize=(12,5))
x= sns.countplot(x='Fruits', hue='Diabetes_012' , data = df_vis);
plt.xlabel('Вживання фруктів')
plt.ylabel('Кількість ')
plt.legend(title='Наявність діабету', labels=['Не має діабету', 'Має діабет', 'Перед діабет'])


x.set_xticklabels(['Так', 'Ні'])
plt.title('Стовпчаста діаграма впливу фактору вживання фруктів на виявлення цукрового діабету ',fontdict={'fontsize':20})
for i in x.patches:
    x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()



#овочі
plt.figure(figsize=(12,5))
x= sns.countplot(x='Veggies', hue='Diabetes_012' , data = df_vis);
plt.xlabel('Вживання овочів')
plt.ylabel('Кількість ')
plt.legend(title='Наявність діабету', labels=['Не має діабету', 'Має діабет', 'Перед діабет'])


x.set_xticklabels(['Так', 'Ні'])
plt.title('Стовпчаста діаграма впливу фактору вживання овочів на виявлення цукрового діабету ',fontdict={'fontsize':20})
for i in x.patches:
    x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()

# HvyAlcoholConsump
df_vis['HvyAlcoholConsump'].value_counts()



#споживання алкоголю
plt.figure(figsize=(12,5))
x= sns.countplot(x='HvyAlcoholConsump', hue='Diabetes_012' , data = df_vis);
plt.xlabel('Алкогольна залежність')
plt.ylabel('Кількість ')
plt.legend(title='Наявність діабету', labels=['Не має діабету', 'Має діабет', 'Перед діабет'])


x.set_xticklabels(['Ні', 'Так'])
plt.title('Стовпчаста діаграма впливу фактору алкогольна залежність на виявлення цукрового діабету ',fontdict={'fontsize':20})
for i in x.patches:
    x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()

plt.figure(figsize=(12,5))
```

```
x= sns.countplot(x='HvyAlcoholConsump', hue='Diabetes_012' , data = df_vis);
for i in x.patches:
    x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()

# Smoker and HvyAlcoholConsump's combined effect on Diabetes
# (1 in Smoker is Yes), (1 in HvyAlcoholConsump is Yes), and (0 is No Diabetes, 1 is Pre Diabetes, 2 is Diabetes)

sns.catplot(x="Smoker" , y ="HvyAlcoholConsump" , data = data , hue="Diabetes_012"  , kind="bar"  );
plt.title("Relation b/w Smoker ,HvyAlcoholConsump and Diabetes")

#HeartDiseaseorAttack
sns.countplot(data=df_vis,x='HeartDiseaseorAttack',hue='Stroke')

plt.figure(figsize=(12,5))

x= sns.countplot(x='HeartDiseaseorAttack', hue='Diabetes_012' , data = df_vis);
for i in x.patches:
    x.annotate('{:.2f}'.format((i.get_height()/df_vis.shape[0])*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()

#The chance of diabetes increases as the person has Heart Disease or Attack
# plt.figure(figsize=(10,6))
sns.countplot(data=df_vis[df_vis['HeartDiseaseorAttack']=="Yes"],x='Stroke',palette='Set1');
#sns.countplot(data=df_vis[df_vis['Diabetes_012']=="2"],x='Stroke',palette='Set1');

#Stroke and HeartDiseaseorAttack's combined effect on Diabetes
# (1 in Stroke is Yes), (1 in HeartDiseaseorAttack is Yes), and (0 is No Diabetes, 1 is Pre Diabetes, 2 is Diabetes)

sns.catplot(x="Stroke" , y ="HeartDiseaseorAttack" , data = data , hue="Diabetes_012"  , kind="bar"  );
plt.title("Relation b/w Stroke ,HeartDiseaseorAttack and Diabetes")

#High blood pressure
plt.figure(figsize=(10,6))
x= sns.countplot(x='Diabetes_012', hue='HighBP' , data = df_vis);
#sns.countplot(data=df_vis,x='Diabetes_012',hue='HighBP',palette='husl')
plt.xlabel('Наявність діабету')
plt.ylabel('Кількість ')
plt.legend(title='АТ', labels=['Так', 'Ні'])


x.set_xticklabels(['Не має діабету', 'Має діабет', 'Перед діабет'])
plt.title('Стовпчаста діаграма впливу фактору АТ на виявлення цукрового діабету ',fontdict={'fontsize':20})

sns.displot(data=df_vis,x='Diabetes_012',col='HighBP',color='#b3b3ff')

#high cholesterol
plt.figure(figsize=(10,6))
x=sns.countplot(data=df_vis,x='Diabetes_012',hue='HighChol',palette='husl')
plt.xlabel('Наявність діабету')
plt.ylabel('Кількість ')
plt.legend(title='Холестерин', labels=['Так', 'Ні'])
```

```
x.set_xticklabels(['Не має діабету', 'Має діабет', 'Перед діабет'])
plt.title('Стовпчаста діаграма впливу фактору холестерину на виявлення цукрового діабету ',fontdict={'fontsize':20})

sns.displot(data=df_vis,x='Diabetes_012',col='HighChol',color='#b3b3ff')

# HighChol with HighBP
plt.figure(figsize=(10,6))
x=sns.countplot(data=df_vis,x='HighChol',hue='HighBP',palette='hls')
for i in x.patches:
    x.annotate('{:.2f}'.format(i.get_height()/df_vis.shape[0]*100)+'%',(i.get_x()+0.25, i.get_height()+0.01))
plt.show()

"""high cholesterol and high blood pressure are highly related to each other as people with high cholesterol tend to have high blood pressure
The link between high blood pressure and high cholesterol goes in both directions. When the body can’t clear cholesterol from the bloodstream, that excess cholesterol can deposit along artery walls. When arteries become stiff and narrow from deposits, the heart has to work overtime to pump blood through them. This causes blood pressure to go up and up.
"""

#Checking HighBP and HighChol's combined effect on Diabetes
# (1 in HighBP is Yes), (1 in HighChol is Yes), and (0 is No Diabetes, 1 is Pre Diabetes, 2 is Diabetes)

sns.catplot(x="HighBP" , y ="HighChol" , data = data , hue="Diabetes_012" , kind="bar" );
plt.title("Relation b/w HighBP ,HighChol and Diabetes")

#BMI
plt.figure(figsize=(12,5))
sns.displot(x='BMI', col='Diabetes_012' , data = df_vis, kind="kde" ,palette="Set2");

sns.set_theme(style="darkgrid")
plt.figure(figsize=(8,6))
fig = sns.scatterplot(data=df_vis, x="Age", y="BMI", hue='Sex')
fig.axhline(y= 25, linewidth=3, color='k', linestyle= '--')
plt.show()

#Split the BMI into (Underweight,Normal weight,Overweight,Obesity)
BMI=pd.cut( data['BMI'],bins=[0,18.5,25,30,80],labels=['Underweight','Normal weight','Overweight','Obesity'])

dd=pd.crosstab(df_vis['Diabetes_012'],BMI,rownames=['Diabetes'])
dd=dd.astype(float)
dd

Diabetes_sum_lst=list(dd.transpose().sum().values)
Diabetes_sum_lst

for idx in range(dd.values.shape[0]):
    dd.values[idx]= dd.values[idx]/Diabetes_sum_lst[idx]*100

dd

dd.plot(kind="bar",figsize=(10,6));
```

```
#sns.countplot(data=df_vis,x='Diabetes_012',hue='HighChol',palette='husl')
plt.xlabel('Наявність діабету')
plt.ylabel('Відсоток ')
plt.legend(title='IMT', labels=['Недостатня вага', 'Нормальна вага', 'Надмірна вага', 'Ожиріння'])

plt.xticks([0, 1, 2], ['Має діабет', 'Не має діабет', 'Перед діабет'], rotation=0)
#x.set_xticklabels(['Має діабету', 'Не має діабет', 'Перед діабет'])
plt.title('Стовпчаста діаграма впливу фактору IMT на виявлення цукрового діабету ',fontdict={'fontsize':20})

plt.figure(figsize=(12,5))
sns.displot(x='Age', col='Diabetes_012' , data = df_vis, kind="kde")
plt.show()



age = pd.cut(df_vis['Age'],bins=[0,4,7,10,12,14],labels=['18:34','35:49','50:64','65:74','75 та старше'])
age

plt.figure(figsize=(10,6))
#plt.title('Стовпчаста діаграма впливу фактору віку на виявлення цукрового діабету ',fontdict={'fontsize':20})
plot=sns.displot(data=df_vis,col='Diabetes_012',x=age,color='#993366');
plot.set_axis_labels("Вік", "Частота")  # Перейменувати вісі

plot.set_titles("Діаграма впливу віку на діабет (якщо не має діабету)")

import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))

g = sns.FacetGrid(data=df_vis, col='Diabetes_012', height=5)
g.map(sns.histplot, 'age', color='#993366')
g.set_axis_labels("Вік", "Частота")

# Перейменувати заголовки для кожного графіку
g.set_titles("Діаграма впливу віку на діабет (якщо не має діабету)", "Діаграма впливу віку на діабет (якщо є діабет)", "Діаграма впливу віку на діабет (якщо є перед діабет)")

plt.show()

#PhysHlth
plt.figure(figsize=(12,5))
sns.displot(x='PhysHlth', col='Diabetes_012' , data = df_vis, kind="kde")
plt.show()

#MentHlth
plt.figure(figsize=(12,5))
x= sns.displot(x='MentHlth', col='Diabetes_012', data = df_vis, kind="kde")
plt.show()

g=sns.displot(data=df_vis.loc[(df_vis['MentHlth']>0)&(df_vis['Diabetes_012']!="No
Diabetes")],x='MentHlth',col='Diabetes_012',col_wrap=2,kde=True);
```

```
new_labels = ["Діабет", "Перед діабет"]
g.set_axis_labels("Психологічне здоровя", "Частота")


plt.show()

#GenHlth
plt.figure(figsize=(10,6))
sns.countplot(data=df_vis,x='Diabetes_012',hue='GenHlth',palette='Set1');
plt.xlabel('Наявність діабету')
plt.ylabel('Кількість ')
plt.legend(title='Оцінка здоровя', labels=['Погана', 'Задовільна','Нормальна', 'Дуже хороша', 'Відмінна'])

plt.xticks([0, 1, 2], ['Не має діабет', 'Має діабет', 'Перед діабет'], rotation=0)
#x.set_xticklabels(['Має діабету', 'Не має діабет', 'Перед діабет'])
plt.title('Стовпчаста діаграма впливу фактору здоровя на виявлення цукрового діабету ',fontdict={'fontsize':20})

plt.figure(figsize=(12,5))
sns.countplot(x='Income', hue='Diabetes_012' , data = df_vis)
plt.show()

# The effect of the income on the healthcare
plt.figure(figsize=(10,6))
sns.displot(data=df_vis,x='Income',col='AnyHealthcare');

plt.figure(figsize=(12,5))
sns.countplot(x='Education', hue='Diabetes_012' , data = df_vis)
plt.show()

#Veggies
pd.crosstab(df_vis.Veggies,df_vis.Diabetes_012).plot(kind="bar",figsize=(5,4))

plt.title('Diabetes Disease Frequency for Veggies')
plt.xlabel("Veggies")
plt.ylabel('Frequency')
plt.show()

#Fruits
pd.crosstab(df_vis.Fruits,df_vis.Diabetes_012).plot(kind="bar",figsize=(5,4))

plt.title('Diabetes Disease Frequency for Fruits')
plt.xlabel("Fruits")
plt.ylabel('Frequency')
plt.show()

#PhysActivity
x=pd.crosstab(df_vis.PhysActivity,df_vis.Diabetes_012).plot(kind="bar",figsize=(5,4))

plt.title('Стовпчаста діаграма впливу фактору фізічної активності на виявлення цукрового діабету')
plt.xlabel("Фізична активність")
plt.ylabel('Частота')
plt.legend(title='Наявність діабету', labels=['Має діабет', 'Не має діабет', 'Перед діабет'])
```

```
x.set_xticklabels(['Hi', 'Так'], rotation =0)
plt.show()

plt.figure(figsize = (10,6))
sns.countplot(data=df_vis,x=df_vis['PhysActivity'],hue='DiffWalk',palette='husl')

plt.figure(figsize = (14,6))
# plt.subplot(1, 1, 1)
x=sns.catplot(data=df_vis[df_vis['BMI']<60],x="PhysActivity", y="BMI", kind="boxen",aspect=1,palette='hls')
plt.show()

y=sns.catplot(data=df_vis[df_vis['BMI']<60],x="DiffWalk", y="BMI", kind="boxen",aspect=1,palette='hls')
plt.show()

plt.figure(figsize=(10,6))
sns.countplot(data=df_vis,x='PhysActivity',hue='GenHlth',palette='Set1');

import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10,6))
sns.countplot(data=df_vis,x='PhysActivity',hue='Diabetes_012',palette='Set1');

#Preprocessing
plt.figure(figsize = (25,8))
u = sns.boxplot(palette = 'cool', data=data)
u.set_xticklabels(u.get_xticklabels(),rotation=45)

data.columns

data.plot(kind="box", subplots=True, layout=(7,4), figsize=(15,14));

#Handling the outliers of the BMI
plt.figure(figsize = (12,6))
plt.subplot(1, 2, 1)
sns.boxplot(data=data,y='BMI',color='#cc6699')
plt.subplot(1, 2, 2)
sns.scatterplot(data=data,x='Diabetes_012',y='BMI',color='#cc6699')
plt.show()

x=data[data['BMI']>=70]
x.shape

df=data.copy()

df=data[data['BMI']<70]

plt.figure(figsize = (12,6))
plt.subplot(1, 2, 1)
sns.boxplot(data=df,y='BMI',color='#cc6699')
plt.subplot(1, 2, 2)
sns.scatterplot(data=df,x='Diabetes_012',y='BMI',color='#cc6699')
```

```
plt.show()

df['Diabetes_012'].value_counts()

#Outlier detection
from sklearn.ensemble import IsolationForest
model = IsolationForest()
model.fit(df)
#df['anomailes_scores']=model.decision_function(df)
df['anomaly']= model.predict(df)

df

#Exclude rows including outliers
df[df['anomaly']==-1]

df[df['anomaly']==-1].shape

df.drop(df[df['anomaly']==-1].index,inplace = True)

df.shape

df

#Drop column(anomaly)
df.drop(columns=['anomaly'], inplace=True)
df.shape

#Scaling Data
x = df.drop(['Diabetes_012'],axis=1)
y = df['Diabetes_012']

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(x)

scaled_features = scaler.transform(x)
x = pd.DataFrame(scaled_features,columns=df.columns[1:])
x.head(10)



#Split the data
x_train , x_test , y_train , y_test = train_test_split(x,y , test_size=0.35, random_state=0, shuffle =True)

#будування
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 5)
knn.fit(x_train,y_train)

y_pred = knn.predict(x_test)
```

```
df['Diabetes_012'].value_counts()

print(classification_report(y_test,y_pred))

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
print(confusion_matrix(y_test,y_pred))

#resampling
from imblearn.combine import SMOTEENN
sm = SMOTEENN()
x_resampled, y_resampled = sm.fit_resample(x,y)

xre_train,xre_test,yre_train,yre_test = train_test_split(x_resampled, y_resampled, test_size=0.3, random_state=42)

knn_smote = KNeighborsClassifier(n_neighbors = 5)
knn_smote.fit(xre_train,yre_train)

yre_pred = knn_smote.predict(xre_test)

print(classification_report(yre_test,yre_pred, labels=[0,1,2]))

df['Diabetes_012'].value_counts()

print(confusion_matrix(y_test,y_pred))

"""Modeling"""

#decision tree
dt= DecisionTreeClassifier(criterion= 'entropy',max_depth=40)
dt.fit(xre_train , yre_train)
print(dt.score(xre_train , yre_train))
print(dt.score(xre_test, yre_test))
y_pred_train_dt = dt.predict(xre_train)
acc_train_dt = accuracy_score(yre_train, y_pred_train_dt)
y_pred_test_dt = dt.predict(xre_test)
acc_test_dt = accuracy_score(yre_test, y_pred_test_dt)
print(acc_train_dt)
print(acc_test_dt)
from sklearn.metrics import classification_report
print(classification_report(yre_test, y_pred_test_dt))
print('Precision: %.3f' % precision_score(yre_test, y_pred_test_dt,average="micro"))
print('Recall: %.3f' % recall_score(yre_test, y_pred_test_dt,average="micro"))
print('F-measure: %.3f' % f1_score(yre_test, y_pred_test_dt,average="micro"))

#Random Forest
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators=100, max_features=16 , max_depth=16)
rf.fit(xre_train,yre_train)
print(rf.score(xre_train, yre_train))
print(rf.score(xre_test, yre_test))
y_pred_train_rf = rf.predict(xre_train)
acc_train_rf = accuracy_score(yre_train, y_pred_train_rf)
```

```
y_pred_test_rf = rf.predict(xre_test)
acc_test_rf = accuracy_score(yre_test, y_pred_test_rf)
print(acc_train_rf)
prit(acc_test_rf)
print(classification_report(yre_test, y_pred_test_rf))
print('Precision: %.3f' % precision_score(yre_test, y_pred_test_rf,average="micro"))
print('Recall: %.3f' % recall_score(yre_test, y_pred_test_rf,average="micro"))
print('F-measure: %.3f' % f1_score(yre_test, y_pred_test_rf,average="micro"))

#Ada Boost
from sklearn.ensemble import AdaBoostClassifier
ada_clf = AdaBoostClassifier( )
ada_clf.fit(xre_train,yre_train)
yhat = ada_clf.predict(xre_test)
from sklearn import metrics
print('Train set Accuracy :',metrics.accuracy_score(yre_train,ada_clf.predict(xre_train))*100)
print('Test set Accuracy :',metrics.accuracy_score(yre_test,yhat)*100)
y_pred_train_ada_clf = ada_clf.predict(xre_train)
acc_train_ada_clf = accuracy_score(yre_train, y_pred_train_ada_clf)
y_pred_test_ada_clf= ada_clf.predict(xre_test)
acc_test_ada_clf = accuracy_score(yre_test, y_pred_test_ada_clf)
print(acc_train_ada_clf)
print(acc_test_ada_clf)
print(classification_report(yre_test, y_pred_test_ada_clf))
print('Precision: %.3f' % precision_score(yre_test, y_pred_test_ada_clf,average="micro"))
print('Recall: %.3f' % recall_score(yre_test, y_pred_test_ada_clf,average="micro"))
print('F-measure: %.3f' % f1_score(yre_test, y_pred_test_ada_clf,average="micro"))
```

Додаток Ж. Сертифікат VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

Рисунок Ж.1 – Сертифікат VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

Додаток З. Тези для VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

АНАЛІЗ ПРИЧИН ТА ПРОГНОЗУВАННЯ ВИЯВЛЕННЯ ЦУКРОВОГО ДІАБЕТУ МЕТОДОМ МАШИННОГО НАВЧАННЯ DECISION TREE

Сидоренко Катерина Віталіївна,
Магістр, Студент,
Національний технічний університет «Дніпровська політехніка»,
Дніпро, Україна

Хом'як Тетяна Валеріївна
к.ф.-м.н., Доцент,
Національний технічний університет «Дніпровська політехніка»,
Дніпро, Україна

Сьогодні кількість людей, які живуть з невиліковними хворобами зростає. Цукровий діабет - це серйозне захворювання, яке може призвести до численних ускладнень та проблем зі здоров'ям. Іноді люди через ускладнення діабету, які не контролюється помирають, а саме може статися інфаркт, гіпоглікемія та інші. Зараз цукровий діабет є однією з найпоширеніших хронічних захворювань у світі, яким страждає близько 530 мільйонів людей, з яких 1 300 000 – громадяни України на червень 2023 року. Це захворювання впливає на рівень цукру (глюкози) в крові. Поява цукрового діабету зазвичай обумовлена генетичними, середовищними та стильовими факторами. Основні причини включають генетичну схильність, ожиріння, неправильну харчову поведінку, інсулінорезистентність та шкідливі звички.

Значущим є той факт, що вчасне виявлення захворювання може запобігти його розвитку. Багато симптомів цукрового діабету, таких як сухість у роті, часті сечовипускання, погіршення зору, втрата ваги, постійне відчуття голоду, не завжди відразу розглядаються як ознаки захворювання. Важливо підкреслити, що ці симптоми можуть бути ранніми показниками високого рівня глюкози у крові.

Отже, для значно більшої ймовірності виявлення цукрового діабету до його появи потрібно мати не тільки більш досвідчених лікарів, а й навчитися прогнозувати дану хворобу для того, щоб у майбутньому не збільшувалась кількість захворюваннях у рік.

Тому потрібно провести аналіз причин, які впливають на ризик розвитку цукрового діабету та зробити прогнозування методом машинного навчання Decision Tree.

Початкові дані для аналізу та прогнозування виявлення цукрового діабету до його появи аби запобігти його розвитку в організмі людини взято з сайту «Центри контролю та профілактики захворювань» [6]. Дані подано у форматі файлу csv, який складаються з 22 колонок (показники та параметри людей) та 253689 рядків (значення показників та параметрів) для кожної людини.

Рисунок 3.1 – Тези для VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

Структуру даних наведено у таблиці 1.

Таблиця 1
Структура даних

Назва колонки	Тип (значення) колонки	Опис колонки
1	2	3
Diabetes_012	Число (0,1,2) 0 – no diabetes 1 – prediabetes 2 – diabetes	Інформація, який описує 0 – пацієнт не має діабету, 1 – пацієнт схильний до діабету (переддіабет), 2 – пацієнт має діабет
HighBP	Число (0,1) 0 – no high BP 1 – high BP	Інформація чи має високий тиск пацієнт: 0 – не має високий тиск, 1 – має високий тиск.
HighChol	Число (0,1) 0 – no high cholesterol 1 – high cholesterol	Інформація про пацієнта чи має високий холестерин: 0 – не має високий холестерин, 1 – має високий холестерин.
CholCheck	Число (0,1) 0 – no cholesterol check in 5 years 1 – yes cholesterol check in 5 yearsCholCheck	Інформація про пацієнти про перевірку холестерину через 5 років: 0 – не було перевірки холестерину через 5 років, 1 – була перевірка холестерину через 5 років.
BMI	Число	Індекс мали тіла
Smoker	Число (0,1) 0 – no 1 – yes	Інформація про пацієнти, чи курив як мінімум 100 цигарок за своє життя: 0 – ні, 1 – так.
Stroke	Число (0,1) 0 – no 1 – yes	Інформація про пацієнта, чи був колись інсульт: 0 – ні, 1 – так.
HeartDiseaseorAttack	Число (0,1) 0 – no 1 – yes	Інформація про пацієнта, чи була колись ішемічна хвороба серця або інфаркт: 0 – ні, 1 – так.
Smoker	Число (0,1) 0 – no 1 – yes	Інформація про пацієнти, чи курив як мінімум 100 цигарок за своє життя: 0 – ні, 1 – так.
Stroke	Число (0,1) 0 – no 1 – yes	Інформація про пацієнта, чи був колись інсульт: 0 – ні, 1 – так.
HeartDiseaseorAttack	Число (0,1) 0 – no 1 – yes	Інформація про пацієнта, чи була колись ішемічна хвороба серця або інфаркт:0 – ні, 1 – так.
PhysActivity	Число (0,1) 0 – no 1 – yes	Інформація про фізичну активність пацієнту протягом останні 30 днів, окрім роботи: 0 – ні, 1 – так.
Fruits	Число (0,1) 0 – no 1 – yes	Інформація про споживання фруктів 1 або більше разів у день пацієнта: 0 – ні, 1 – так.

Рисунок 3.2 – Тези для VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

Продовження табл. 1

1	2	3
Veggies	Число (0,1) 0 – no 1 – yes	Інформація про споживання овочів 1 або більше разів у день пацієнта: 0 – ні, 1 – так.
HvyAlcoholConsump	Число (0,1) 0 – no 1 – yes	Інформація про споживання алкоголю (для чоловіків більше 14 напоїв на тиждень, для жінок – 7): 0 – ні, 1 – так.
AnyHealthcare	Число (0,1) 0 – no 1 – yes	Інформація про наявність медичного страхування у пацієнта: 0 – ні, 1 – так.
NoDocbcCost	Число (0,1) 0 – no 1 – yes	Інформація про похід пацієнта до будь-якого лікаря протягом останніх 12 місяців: 0 – ні, 1 – так.
GenHlth	Число (1-5) 1 – excellent, 2 – very good, 3 – good, 4 – fair, 5 – poor.	Інформація про здоров'я пацієнту на його думку за шкалою: 1 – відмінна, 2 – дуже хороша, 3 – нормальна, 4 – задовільна, 5 – погана.
MenthHlth	Число (1-30)	Інформація про психічне здоров'я, де пацієнт мав стрес, депресію та проблеми з емоціями протягом останніх 30днів. Треба записати скільки днів пацієнт мав такий стан від 1 до 30.
PhysHlth	Число (1-30)	Інформація про фізичне здоров'я, де пацієнт мав фізичні захворювання та травми протягом останніх 30днів. Треба записати скільки днів пацієнт мав такий стан від 1 до 30.
DiffWalk	Число (0,1) 0 – no 1 – yes	Інформація про пацієнта чи є серйозні труднощі при ходьбі чи підйомі по сходах: 0 – ні, 1 – так.
Sex	Число (0,1) 0 – female 1 - male	Стать пацієнта: 0 – жінка, 1 – чоловік.
Age	Число	Вік, де поділено на 13 категорій: 1 – 18-24, 13 – 80 і більше років.
Education	Число (1-6) 1 – never attended school 2 – elementary grades 3 – some high school graduate 4 – high school graduate 5 – college year to 3 years 6 – college 4 years or more	Інформація про освіту пацієнта: 1 – ніколи не відвідував школу, 2 – 1-8 класи, 3 – 9-11 класи, 4 – 12 класів, 5 – коледж до 3 років, 6 – закінчив коледж.

Рисунок 3.3 – Тези для VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

Продовження табл. 1

1	2	3
Income	Число (1-8)	Інформація про заробітну плату, яка розподілена на 8 груп, де 1 – менше ніж 10000$, 8 – більше, ніж 75000$.

Набір даних у даному дата сеті є незбалансованим, оскільки значення 0 (не має діабету) зустрічається 213703 рази, 1 (перед діабет) зустрічається 35346 раз та 2 (має діабет) зустрічається 4631 раз. Цей фактор потрібно врахувати в процесі попереднього опрацювання даних.

Основні кроки попереднього опрацювання даних включають наступне:

–збір та завантаження даних;

–очищення даних (Data Cleaning): спочатку було видалено дані, які мали дублікати, кількість знайдених дублікатів дорівнює 15548 та даних зі значенням NaN або NULL не було знайдено у даному дата сеті;

–візуалізація даних (Data Visualization): аналіз та візуалізація даних, щоб отримати більше інсайтів про їх розподіл та взаємозв'язки між ознаками;

–видалення викидів (outliers) - це важлива складова попереднього опрацювання даних, і це важливо зробити, коли у ваших даних є значення, які суттєво відрізняються від інших і можуть спотворити результати аналізу або моделі. Видалення викидів за допомогою методу Isolation Forest [8], де було знайдено та видалено 114071 аномальних значень, тобто після видалення аномалій дата сет має 114071 рядків значень показників пацієнтів;

–відбір ознак (Feature Selection): для даного набору даних при першому аналізі кожної колонки даних було виявлено, що колонки Education, Income, AnyHealthcare, є неінформативними для виявлення цукрового діабету у людини;

–розділення даних (Data Splitting): для даного набору даних було обрано, що 70% даних – навчальна вибірка та 30% – тестова;

–балансування класів (Class Balancing): для даного набору даних було обрано метод SMOTEENN [8]. Після застосування метода SMOTEENN дані були збалансовані, де класи мали наступні значення: 0 – 27948 значення, 1 – 39597, 2 – 35628 значень.

– масштабування атрибутів: для даного набору даних було обрано метод Standard Scaler [8].

Проаналізувавши діаграму впливу факторів на цукровий діабет, яка зображена на рис. 1, можна зробити висновок, що всі фактори впливу на виявлення діабету є слабко корельовані. Але то не значить, що існує відсутність зв'язку між змінними, а вказує на то, що зв'язок обмежений та має невеликий вплив.

Рисунок 3.4 – Тези для VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

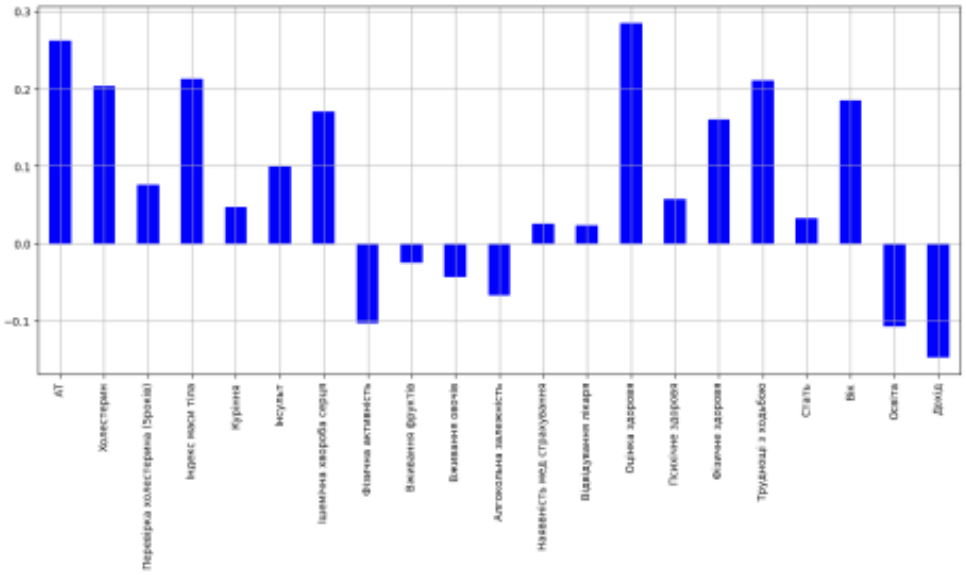


Рисунок 1 Діаграма кореляції факторів, які впливають на виявлення діабет

Після детального аналізу кожного показнику, який на виявлення діабету та зробили наступні висновки:

–пацієнти, які мають ішемічну хворобу серця (ІХС), хворіють приблизно у 4 рази більше на цукровий діабет;

–лише 27% пацієнтів курять, а з них 8% мають цукровий діабет або перед діабет;

–жінки-пацієнти, які не мають діабету, на 20000 більше, ніж чоловіки. Оскільки, жінки протягом життя мають різні гормональні зміни (вагітність, менопауза та інші);

–пацієнти, які не вживають фрукти, майже у 2 рази більше не хворіють на цукровий діабет. Свідчить про то, що не всі фрукти можуть нести корисні речовини та деякі з них містять багато цукру;

–овочі можуть бути важливим фактором для профілактики діабету, але важливо дотримуватися збалансованого споживання;

–пацієнти з високим холестерином мають у два рази частіше цукровий діабет. Оскільки, високий АТ може сприяти пошкодженню судин, погіршувати контроль рівня цукру в крові й підвищувати ризик серцево-судинних захворювань;

–пацієнти у віці від 50 до 64 років складають найбільшу кількість тих, хто має діабет. Оскільки, відбувається нормальні процеси старіння, зниження функції підшлункової залози та збільшення інсулінорезистентності;

–кількість пацієнтів з ожиріння та цукровим діабетом складає 50%;

–частина пацієнтів, які мають інсульт та діабет, складає лише 1,5%;

Рисунок 3.5 – Тези для VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

–фізично активні пацієнти мають діабет у 3,5 раза менше, ніж інші. Це підкреслює важливість фізичної активності для зменшення ризику розвитку діабету.

Наступним кроком було здійснено прогнозування методом Decision Tree, де модель (Decision Tree Classifier) мала такі гіперпараметри: criterion= 'entropy', max_depth=40, а інші параметри за замовченням та отримано результати, які наведені у таблиці 2.

Таблиця 2
Результати моделі Decision Tree Classifier

class	precision	recall	f1-score	accuracy
0(не має діабет)	0.92	0.89	0.91	0.92
1(переддіабет)	0.93	0.95	0.94	
2 (має діабет)	0.89	0.89	0.89	

З таблиці 2 можна зробити висновки, що точність моделі для всіх класів дорівнює 0.92, що свідчить про її загальну ефективність у класифікації. Модель має високі показники precision та recall для класів "Не має діабету" і "Пред діабет" що робить її корисною для виявлення цих станів у пацієнтів. Однак для класу "Має діабет" її ефективність менша, але все одно прийнятна.

Отримані результати дослідження мають велике значення, оскільки вони можуть бути використані для покращення роботи медичних фахівців у виявленні цукрового діабету до його появи аби запобігти його розвитку в організмі людини та підвищити шанси порятунку життя та здоров'я пацієнтів, які стикаються з цією серйозною та невиліковною хворобою. Результати роботи також можуть бути корисними для розробки та моделювання інформаційних систем з прогнозу виявлення цукрового діабету [9].

Список літератури:

1. Машинне навчання простими словами. Частина 1. URL: http://www.mmf.lnu.edu.ua/ar/1739 (дата звернення: 01.10.2023 року)

2. Оцінка якості моделі класифікації. URL: https://studfile.net/preview/9974842/page:22/ (дата звернення: 05.10.2023 року)

3. Кількість діабетиків у світі до 2050 року може зрости майже втричі. URL: https://thepage.ua/ua/news/kilkist-diabetikiv-v-sviti-mozhe-zrosti-do-13-milyarda-lyudej-do-2050-roku (дата звернення: 02.10.2023 року)

4. Outliers in Machine Learning. URL: https://medium.com/analytics-vidhya/outliers-in-machine-learning-e830b2bd8660 (дата звернення: 01.10.2023 року)

5. Isolated Forest. URL: https://medium.com/@corymaklin/isolation-forest-799fceacdda4 (дата звернення: 03.10.2023 року)

Рисунок 3.6 – Тези для VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»

6. Standard Scaler. URL: learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (дата звернення: 03.10.2023 року)
7. Parameters and Hyperperemeters in Machine Learning and Deep Learning. URL: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac (дата звернення: 05.10.2023 року)
8. Decision Tree Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (дата звернення: 05.10.2023 року)
9. Ус С.А., Слєсарєв В.В., Хом'як Т.В., Козир С.В. Моделювання та реінжиніринг бізнес-процесів: Навчальний посібник / Дніпро: НТУ «Дніпровська політехніка», 2020. – 180 с.

Рисунок 3.7 – Тези для VI International Scientific and Practical Conference «METHODICAL AND PRACTICAL METHODS OF CREATING INVENTIONS»