Міністерство освіти і науки України
Харківський національний університет імені В. Н. Каразіна
Навчально-науковий інститут комп'ютерних наук та штучного інтелекту
Кафедра комп'ютерних систем та робототехніки

До захисту допущено
Кафедрою комп'ютерних систем та робототехніки
протокол № __ від __ грудня 2025р.

завідувач кафедри _____ Максим ХРУСЛОВ
(підпис)

«___»_____ 2025 р.

Кваліфікаційна робота
здобувача першого (бакалаврського) рівня вищої освіти

«Комп'ютерна система класифікації станів медико-біологічних систем за допомогою методу випадкових лісів»

Спеціальність ***123 – Комп'ютерна інженерія.***
Освітня програма ***Комп'ютерна інженерія***

Виконавець _____ **Артем НІКОЛАЙЧУК**
(підпис)

Науковий керівник _____ **Ніна БАКУМЕНКО**
(підпис)

АНОТАЦІЯ

Пояснювальна записка до кваліфікаційної роботи бакалавра складається зі вступу, трьох розділів, висновків, списку використаних джерел і трьох додатків. Загальний обсяг роботи складає 63 сторінки, із яких 47 сторінок основної частини з 10 рисунками, 2 таблицями, 33 найменуваннями списку використаних джерел та трьома додатками.

Об'єкт дослідження – процес прогнозування стадії захворювання на цукровий діабет за допомогою аналізу клінічних показників крові.

Предмет дослідження – методи та моделі машинного навчання, орієнтовані на підвищення точності визначення стадії цукрового діабету.

Мета дослідження – підвищення ефективності прогнозування стадії захворювання на цукровий діабет шляхом розроблення та реалізації моделі, що базується на методі випадкових лісів і дозволяє інтерпретувати отримані результати для медичних спеціалістів.

Методи дослідження – структурний і системний аналіз для формалізації предметної області; методи машинного навчання для побудови моделей класифікації; методи математичної статистики для оцінки точності; пояснювальні моделі (SHAP) для аналізу важливості ознак.

Наукова новизна роботи полягає у створенні пояснювальної моделі прогнозування стадії цукрового діабету на основі методу випадкових лісів, яка дає змогу не лише підвищити точність класифікації, а й визначити найбільш інформативні клінічні показники, що впливають на прогноз.

Практичне значення роботи полягає у розробленні програмного застосунку для автоматизованого прогнозування стадії цукрового діабету за біохімічними параметрами крові, що може бути використаний у медичних інформаційних системах для підтримки прийняття діагностичних рішень.

Ключові слова: машинне навчання, випадковий ліс, класифікація даних, цукровий діабет, медико-біологічні системи, клінічні показники крові, прогнозування стадії захворювання, пояснюваний штучний інтелект, SHAP-аналіз, медичні інформаційні системи, підтримка прийняття рішень, комп'ютерна система.

ABSTRACT

An explanatory note to the master's attestation work is created in the introduction, three sections, conclusions, a list of sources used and three additional substances.

The total volume of work is 63 pages, of which 50 pages of the main part with 10 figures, 2 table, 33 names of the list of used sources and two additions.

Object of the research – the process of predicting the stage of diabetes mellitus through the analysis of clinical blood indicators.

Subject of the research – machine learning methods and models aimed at improving the accuracy of diabetes stage determination, particularly using the random forest algorithm and SHAP explainability techniques.

Aim of the research – to increase the efficiency of predicting the stage of diabetes mellitus by developing and implementing a model based on the random forest method that allows the interpretation of the obtained results for medical specialists.

Research methods – structural and system analysis for the formalization of the problem domain; machine learning methods for building classification models; mathematical statistics methods for accuracy evaluation; and explainable models (SHAP) for feature importance analysis.

Scientific novelty of the work lies in the development of an explainable model for predicting the stage of diabetes mellitus based on the random forest method, which not only improves classification accuracy but also identifies the most informative clinical indicators influencing the prediction.

Practical significance of the work consists in the development of a software application for automated prediction of the diabetes stage based on biochemical blood parameters, which can be integrated into medical information systems to support diagnostic decision-making.

Keywords: machine learning, random forest, data classification, diabetes mellitus, medico-biological systems, clinical blood indicators, disease stage prediction, explainable artificial intelligence, SHAP analysis, medical information systems, decision support systems, computer system.

ЗМІСТ

ВСТУП

Сучасна медицина приділяє особливу увагу удосконаленню методів діагностики та прогнозування різноманітних захворювань, оскільки точне визначення стану пацієнта є критично важливим для забезпечення ефективного лікування та запобігання розвитку ускладнень. У багатьох випадках успішна терапія безпосередньо залежить від правильного та своєчасного встановлення діагнозу, що дозволяє адаптувати лікувальні стратегії з урахуванням індивідуальних особливостей організму пацієнта та характеру захворювання [1]. У цьому контексті розробка комп'ютеризованих систем, здатних автоматизувати процес аналізу медичних даних та надавати підтримку у прийнятті клінічних рішень, набуває особливої актуальності, оскільки забезпечує підвищення об'єктивності оцінки стану пацієнтів і зменшує ймовірність людських помилок.

Актуальність дослідження визначається постійним зростанням обсягів медичної інформації, яка формується у процесі ведення електронних медичних записів, лабораторних досліджень та інших клінічних джерел. Традиційні підходи до діагностики часто мають обмеження у точності та ефективності, потребують значних часових і людських ресурсів, а також не завжди здатні одночасно інтегрувати різні типи інформації для отримання комплексного уявлення про стан пацієнта. У зв'язку з цим особливе значення набуває використання алгоритмів машинного навчання, які дозволяють обробляти великі масиви даних, виявляти приховані закономірності та забезпечувати високоточне прогнозування медичних станів, що істотно підвищує ефективність прийняття клінічних рішень [2].

Мета дослідження полягає у розробці комп'ютеризованої системи для автоматизованого визначення медичних діагнозів на основі аналізу клінічних показників пацієнтів із використанням алгоритму Random Forest, що забезпечує високу точність класифікації.

Об’єкт дослідження – процес автоматизованого аналізу медичних даних, що дозволяє встановлювати діагноз пацієнта на основі комплексного обліку клінічних параметрів та попередньо накопиченої медичної інформації.

Предмет дослідження – методи машинного навчання, зокрема алгоритм Random Forest, які застосовуються для класифікації та прогнозування медичних станів на основі числових клінічних даних.

Практичне значення отриманих результатів полягає у можливості інтеграції розробленої комп’ютеризованої системи у сучасні медичні інформаційні платформи та її застосування для підтримки прийняття рішень лікарями. Використання запропонованого підходу дозволяє підвищити точність і швидкість діагностики, зменшити навантаження на медичних фахівців та забезпечити науково обґрунтовану основу для подальшого розвитку автоматизованих систем підтримки клінічних рішень у різних галузях медицини. Крім того, реалізація даної системи створює умови для стандартизації процесів обробки клінічних даних та забезпечує більш ефективний контроль якості надання медичної допомоги пацієнтам.

РОЗДІЛ 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Огляд інформаційних систем та технологій, що використовуються для діагностики та прогнозування результатів медичних аналізів

Сучасний розвиток медичної інформатики характеризується широким впровадженням інформаційних систем та технологій, орієнтованих на автоматизацію процесів діагностики і прогнозування результатів медичних аналізів. В умовах зростання обсягів клінічних даних та підвищення складності діагностичних процедур виникає потреба у створенні комплексних програмно-апаратних рішень, здатних інтегрувати лабораторні результати, клінічні спостереження та історії хвороби в єдине інформаційне середовище. Такі системи спрямовані не лише на зберігання й обробку даних, але й на побудову інтелектуальних моделей, що забезпечують лікаря підтримкою у прийнятті рішень [3].

Важливим аспектом розвитку інформаційних технологій у медицині є використання електронних медичних записів, які забезпечують централізований доступ до результатів аналізів, історії хвороб та супутніх діагностичних даних. Поєднання електронних записів із системами управління лабораторними процесами створює основу для побудови прогнозних моделей, які ґрунтуються на статистичному аналізі та машинному навчанні. У цьому контексті особливого значення набувають методи глибинного навчання, що дозволяють виявляти приховані залежності у великих масивах даних та формувати прогностичні висновки з високою точністю.

Розвиток технологій штучного інтелекту сприяє формуванню нових підходів до інтерпретації медичних аналізів. Використання алгоритмів класифікації, регресії та кластеризації уможливлює автоматичне визначення патологічних відхилень, а також створення індивідуалізованих прогнозів перебігу захворювань. Такі підходи забезпечують підвищення ефективності

клінічних досліджень, скорочення часу на аналіз результатів та мінімізацію людського фактору.

Значний розвиток отримали хмарні технології, які дозволяють інтегрувати інформаційні системи різних медичних закладів у єдиний аналітичний простір. Вони забезпечують масштабованість, швидкий доступ до даних та можливість їх обробки у реальному часі. Хмарні рішення відкривають перспективи для створення розподілених систем прогнозування, які використовують накопичений досвід багатьох медичних установ та дозволяють отримати більш обґрунтовані висновки щодо стану пацієнтів.

Важливим напрямом є також застосування технологій великих даних. Аналіз багатовимірних структурованих та неструктурованих даних медичного характеру потребує високопродуктивних алгоритмів, здатних працювати з мільйонами записів одночасно [4]. Використання таких технологій робить можливим побудову складних діагностичних моделей, що враховують широкий спектр біохімічних показників, генетичну інформацію та дані візуальної діагностики.

1.2 Огляд інформаційних систем та технологій, що використовуються для діагностики та прогнозування результатів медичних (аналізів)

Медичні інформаційні системи, що функціонують на основі алгоритмів машинного навчання, становлять сучасний інструмент підтримки клінічних рішень, автоматизації діагностичного процесу та прогнозування перебігу захворювань. Їх застосування зумовлене потребою у комплексному аналізі великих обсягів даних, до складу яких входять електронні медичні записи, результати лабораторних обстежень, інформація про генетичні фактори, показники способу життя пацієнтів та їхній клінічний анамнез. Залучення методів машинного навчання забезпечує виявлення прихованих закономірностей, багатофакторних залежностей та формування індивідуалізованих прогнозних моделей.

Архітектура зазначених систем є багаторівневою і включає такі ключові компоненти [5]:

– модулі збору та попередньої обробки даних, що забезпечують акумулювання інформації з лабораторних комплексів, сенсорів безперервного моніторингу фізіологічних параметрів, мобільних застосунків та натільних пристроїв;

– аналітичні алгоритмічні блоки, побудовані на базі методів машинного навчання, які виконують класифікацію клінічних станів, прогнозування ризиків та моделювання ймовірнісних сценаріїв перебігу патологічних процесів;

– користувацькі інтерфейси, що слугують засобом інтеграції результатів діагностики та прогнозування у практику лікаря, а також забезпечують пацієнтам доступ до персоналізованих рекомендацій і можливість відстеження стану здоров’я у динаміці.

Використання таких систем сприяє трансформації традиційних підходів до медичної практики у напрямі персоналізованої медицини, орієнтованої на індивідуальні характеристики пацієнта. Це дозволяє підвищити точність діагностики, своєчасно виявляти патологічні зміни та оптимізувати терапевтичні стратегії. Водночас їхня ефективність залежить від якості даних, рівня алгоритмічної інтеграції та відповідності вимогам безпеки і конфіденційності медичної інформації.

1.2.1 Медична інформаційна система MedAI

Система MedAI представляє собою інтегроване програмно-аналітичне рішення для автоматизованого оцінювання ризиків та прогнозування клінічних станів пацієнтів на основі методів машинного навчання. Функціонування системи забезпечується застосуванням алгоритмів ансамблевого навчання, зокрема Random Forest, які формують сукупність дерев рішень для кількісного визначення ймовірності розвитку патологічних процесів [6].

На вхід системи вводяться параметри пацієнта, що включають демографічні характеристики, фізіологічні показники, дані про спосіб життя, клінічний анамнез та результати лабораторних досліджень. Вхідна інформація піддається порівняльному аналізу з наявними історичними медичними даними та клінічними дослідженнями, що дозволяє алгоритму ідентифікувати найбільш значущі фактори ризику та формувати прогноз у вигляді ймовірнісної оцінки розвитку патології.

Архітектура системи MedAI є модульною і включає наступні ключові компоненти:

– модуль збору та попередньої обробки даних, який забезпечує валідацію вхідних параметрів, заповнення пропущених значень статистичними методами та нормалізацію даних для забезпечення коректної роботи алгоритмів машинного навчання;

– модуль машинного навчання, що реалізує алгоритми ансамблевого навчання для оцінки ризиків та методи класифікації, такі як опорні вектори (SVM), для визначення стану пацієнта та прогнозування розвитку патологічних процесів;

– модуль обробки історичних медичних даних, який аналізує клінічні записи та бази досліджень з метою виявлення аналогічних випадків та підвищення точності прогнозу;

– модуль бази даних (Data Storage Layer), що забезпечує збереження медичних та навчальних даних, включаючи анонімізовані клінічні записи та результати лабораторних досліджень, необхідні для навчання та оновлення алгоритмів;

– модуль інтерфейсу користувача (UI/UX), що забезпечує лікарям доступ до введення даних, перегляду прогнозних показників у вигляді ймовірності розвитку патології та отримання персоналізованих рекомендацій;

– модуль автентифікації та безпеки, який реалізує контроль доступу за ролями користувачів та шифрування даних, гарантує захист конфіденційної

медичної інформації під час її зберігання та передавання між компонентами системи.

Модульна структура системи MedAI дозволяє ефективно застосовувати алгоритми машинного навчання для автоматизованого прогнозування клінічних станів, інтегрувати систему з існуючими медичними інформаційними платформами та забезпечувати високий рівень захисту даних. Застосування MedAI сприяє підвищенню точності оцінки ризиків, автоматизації процесів класифікації та прогнозування, а також забезпеченню лікарів науково обґрунтованими інструментами для прийняття рішень.

1.2.2 Медична інформаційна система PandaCare ML

Система PandaCare ML являє собою модульне програмно-аналітичне рішення для комплексного аналізу медичних даних та прогнозування клінічних станів пацієнтів на основі методів машинного навчання. У її функціонуванні застосовується алгоритм XGBoost (Extreme Gradient Boosting), який реалізує градієнтний бустинг і забезпечує високу точність прогнозування шляхом комбінування численних слабких моделей, представлених деревами рішень [7].

Система приймає на вхід широкий спектр клінічних та біомедичних показників, включаючи лабораторні дані, біомаркери, генетичні предиктори та інші клінічно релевантні фактори. Алгоритм XGBoost аналізує взаємозв’язки між цими показниками та формує прогностичну модель, здатну оцінювати ймовірність розвитку певних станів з високою точністю.

Для забезпечення пояснюваності результатів у PandaCare ML додатково застосовується метод дерева рішень (Decision Tree). На відміну від складних «чорних скриньок» нейронних мереж, використання дерев рішень дозволяє відстежувати вплив окремих факторів на сформований прогноз та обґрунтування, на підставі якого система прийняла відповідні висновки.

Архітектура системи побудована за модульним принципом та включає такі основні компоненти:

– модуль інтерфейсу користувача (UI/UX), що забезпечує лікарям доступ до введення медичних даних, перегляду результатів прогнозування та отримання пояснень щодо впливу окремих факторів. Модуль містить інтерактивні візуалізації для демонстрації взаємозв’язків між показниками та їхнього впливу на прогностичну модель;

– модуль бази даних (Data Storage Layer), який містить історичні клінічні записи, результати лабораторних досліджень та дані клінічних досліджень. Модуль підтримує оновлення навчальних наборів даних і використовує механізми кешування для оптимізації обробки запитів;

– модуль безпеки та контролю доступу, що реалізує автентифікацію користувачів, шифрування даних та контроль доступу відповідно до ролей, забезпечуючи захист конфіденційної медичної інформації та відповідність стандартам безпеки.

Модульна архітектура PandaCare ML забезпечує ефективну інтеграцію алгоритмів машинного навчання з медичними інформаційними платформами та дозволяє отримувати точні прогностичні оцінки з високим рівнем пояснюваності. Використання системи сприяє підвищенню точності оцінки ризиків, автоматизації аналітичних процесів та наданню лікарям обґрунтованих інструментів для прийняття клінічних рішень.

1.2.3 Медична інформаційна система DeepHealth

Система DeepHealth спеціалізується на комплексній обробці медичних зображень та аналізі часових рядів даних, що надходять від безперервного моніторингу фізіологічних параметрів пацієнтів. Для обробки зображень застосовуються згорткові нейромережі (CNN), які дозволяють автоматично ідентифікувати патологічні зміни та класифікувати їх за рівнем складності [8]. Це забезпечує раннє виявлення аномалій та підтримку клінічних рішень.

Для аналізу та прогнозування динаміки фізіологічних показників використовується архітектура рекурентних нейромереж із довготривалою

короткочасною пам’яттю (Long Short-Term Memory, LSTM). Моделі LSTM обробляють часові ряди даних, враховуючи історичні тренди та зовнішні фактори, що дозволяє прогнозувати подальші зміни параметрів пацієнта. Такий підхід підвищує точність прогнозів і підтримує прийняття клінічних рішень на основі динамічного аналізу даних.

Архітектура системи DeepHealth побудована за модульним принципом та включає наступні компоненти:

– модуль інтерфейсу користувача (UI/UX), який забезпечує можливість завантаження медичних зображень, перегляду результатів аналізу та прогнозів динаміки фізіологічних параметрів. Модуль включає інтерактивні візуалізації для представлення змін показників у часі та факторів, що на них впливають;

– модуль обробки зображень, що використовує CNN для аналізу медичних знімків, ідентифікації аномалій та класифікації їх за ступенем вираженості;

– модуль аналізу часових рядів, який застосовує LSTM для прогнозування подальшого розвитку фізіологічних параметрів на основі історичних даних та додаткових клінічних характеристик;

– модуль зберігання даних, що забезпечує архівування медичних зображень, часових рядів та клінічних параметрів пацієнтів. Для оптимізації доступу та зберігання великих обсягів даних застосовуються розподілені файлові системи, об’єктні сховища та бази даних часових рядів, що дозволяє ефективно управляти різнотипною інформацією;

– модуль безпеки та контролю доступу, який реалізує автентифікацію користувачів, шифрування даних та обмеження доступу до чутливої інформації відповідно до ролей користувачів.

Згорткові моделі у системи DeepHealth забезпечують високу точність розпізнавання патернів на медичних зображеннях, тоді як рекурентні компоненти ефективно обробляють часові ряди даних. Така інтеграція дозволяє системі одночасно виконувати візуальну діагностику та прогнозувати динаміку фізіологічних показників пацієнта.

1.2.4 Порівняльний аналіз розглянутих медичних інформаційних систем

Медичні інформаційні системи характеризуються спеціалізацією на різних напрямках аналізу: частина з них орієнтована на ідентифікацію факторів ризику та раннє виявлення патологічних станів, тоді як інші спрямовані на прогнозування потенційних ускладнень або безперервний моніторинг динаміки стану пацієнта.

У табл. 1.1 показано порівняльний аналіз медичних інформаційних систем MedAI, PandaCare ML та DeepHealth.

Таблиця 1.1

Порівняльний аналіз медичних інформаційних систем

Параметр	MedAI	PandaCare ML	DeepHealth
Основні алгоритми обробки даних	Random Forest та SVM для класифікації клінічних показників	XGBoost та Decision Tree для прогнозування та пояснюваності	Згорткові нейронні мережі (CNN) та моделі LSTM для аналізу зображень і часових рядів
Основне функціональне призначення	Автоматизована оцінка ймовірності розвитку захворювання	Комплексний аналіз ризиків та прогнозування стану пацієнта	Виявлення патологічних змін у медичних зображеннях та прогнозування динаміки клінічних параметрів
Типи вхідних даних	Клінічні та лабораторні показники, антропометричні дані, анамнез	Лабораторні показники, генетичні маркери, інформація про спосіб життя	Медичні зображення (рентген, офтальмологічні знімки), часові ряди лабораторних даних
Можливості прогнозування ризику	Присутні, оцінка ймовірності розвитку захворювання	Присутні, із врахуванням багатьох факторів ризику	Відсутні, фокус на аналізі змін та ускладнень
Можливості діагностики	Первинний скринінг стану пацієнта	Клінічна діагностика на основі	Обмежено, лише оцінка ускладнень і

		багатовимірного аналізу даних	динаміки параметрів

Кінець таблиці 1.1

Параметр	MedAI	PandaCare ML	DeepHealth
Прогнозування ускладнень	Не реалізоване	Частково, через оцінку ризиків	Присутнє, аналіз патологічних змін та динаміки показників
Інтерпретованість результатів	Відносно висока, пояснення через структуру дерев рішень	Висока, прозора інтерпретація факторів ризику	Обмежена, потребує додаткової експертної оцінки через складність нейромереж
Продуктивність обробки даних	Висока швидкість обробки	Середня швидкість обробки	Середня швидкість обробки
Обмеження та недоліки	Обмежена глибина аналізу складних даних, залежність від якості вхідних показників	Потребує великих обсягів даних для навчання, деякі алгоритми менш оптимальні за швидкістю	Високі обчислювальні ресурси, складність інтеграції з іншими медичними системами, потребує перевірки експертами

Сучасні медичні інформаційні системи демонструють значний аналітичний потенціал, проте кожна з них спеціалізується на окремих аспектах обробки даних. MedAI застосовує Random Forest та SVM для оцінки ризиків на основі лабораторних та антропометричних показників, виконуючи первинний скринінг, але обмежена у глибині аналізу. PandaCare ML інтегрує XGBoost та Decision Tree, забезпечуючи комплексний аналіз ризиків та пояснюваність результатів, однак потребує великих обсягів даних та обчислювальних ресурсів. DeepHealth орієнтована на обробку медичних зображень і часових рядів, використовуючи CNN та LSTM, проте її результати складно інтерпретувати, а інтеграція з іншими системами обмежена.

Порівняльний аналіз обґрунтовує необхідність розробки власної комп’ютеризованої системи, здатної поєднувати аналітичні можливості різних

рішень, забезпечувати ефективну обробку даних та інтерпретованість результатів для підвищення якості діагностики та прийняття клінічних рішень.

1.3 Постановка задачі дослідження

Мета дослідження полягає у розробці комп’ютеризованої системи для автоматизованого визначення медичних діагнозів на основі аналізу клінічних показників пацієнтів із використанням алгоритму Random Forest, що забезпечує високу точність класифікації.

Задля досягнення мети треба вирішити такі завдання:

– розглянути ключові клінічні показники, які використовуються для оцінки стану пацієнта;

– виконати аналіз існуючих методів прогнозування стану пацієнта на основі біохімічних показників крові;

– розробити метод прогнозування стану пацієнта на основі алгоритму випадкових лісів;

– визначити та описати вхідні дані для системи прогнозування, включно з різнорідними клінічними параметрами;

– створити програмну реалізацію комп’ютеризованої системи прогнозування стану пацієнта з використанням випадкових лісів;

– провести експериментальне дослідження для перевірки ефективності запропонованого методу;

– оцінити точність прогнозування та порівняти результати розробленого підходу з іншими методами машинного навчання.

РОЗДІЛ 2 РОЗРОБЛЕННЯ МОДЕЛІ ВИЗНАЧЕННЯ СТАДІЙ ЗАХВОРЮВАННЯ ЦУКРОВОГО ДІАБЕТУ ЗА КЛІНІЧНИМИ ПОКАЗНИКАМИ КРОВІ

2.1 Огляд методів машинного навчання для визначення стадії захворювання цукрового діабету

2.1.1 Метод баєсівського класифікатора

Баєсівський класифікатор ґрунтується на застосуванні теореми Байєса та функціонує за принципом обчислення ймовірностей належності спостереження до певного класу. Сутність методу полягає в оцінюванні апостеріорної ймовірності кожного класу на основі заданих ознак, після чого обирається клас із найвищим значенням цієї ймовірності [9].

У завданнях прогнозування цукрового діабету цей підхід дає змогу визначити ймовірність виникнення або прогресування захворювання, спираючись на медико-біологічні показники пацієнта. На початковому етапі формується навчальна вибірка, яка містить значення ознак (вік, індекс маси тіла, рівень глюкози в крові, артеріальний тиск, інсулінорезистентність тощо) та відповідні мітки класів, що характеризують стан пацієнта (наприклад, «високий ризик діабету» або «низький ризик»). Ці дані використовуються для оцінювання ймовірностей, необхідних для подальшої класифікації нових випадків.

Далі алгоритм обчислює апріорні ймовірності для кожного класу, що дозволяє визначити базову ймовірність належності пацієнта до певної категорії ще до врахування конкретних медичних ознак. Паралельно визначаються умовні ймовірності появи кожної ознаки в межах кожного класу.

Під час класифікації нового пацієнта алгоритм обчислює ймовірність того, що спостереження з конкретними характеристиками належить до кожного з можливих класів. Таким чином, на основі комбінації показників (наприклад, підвищеного рівня глюкози чи надлишкової маси тіла) визначається ступінь

ризику розвитку цукрового діабету. Остаточне рішення приймається шляхом порівняння отриманих ймовірностей – пацієнт відноситься до того класу, для якого апостеріорна ймовірність є максимальною.

Основними перевагами баєсівського класифікатора є простота реалізації, ефективність при достатньому обсязі навчальних даних і стійкість до пропусків у наборі ознак [10]. Крім того, метод може працювати з різнотипними медичними показниками, що робить його придатним для задач медичного прогнозування. Водночас основним обмеженням залишається припущення про статистичну незалежність ознак, яке не завжди відповідає реальним умовам, оскільки між медичними параметрами часто існують суттєві кореляції.

Формально, для кожного зразка $x = (x_1, x_2, ..., x_d)$ ймовірність приналежності до класу y_i визначається таким способом:

$$P(y_j|x) = \frac{P(x|y_j)P(y_j)}{P(x)}, \tag{2.1}$$

де $P(y_j|x)$ – апостеріорна ймовірність належності об’єкта до y_i класу;

$P(x|y_i)$ – умовна ймовірність ознак за умови, що зразок належить до y_i класу;

$P(y_j)$ – апріорна ймовірність появи класу y_i;

$P(x)$ – загальна ймовірність спостереження ознак.

Метод ґрунтується на припущенні статистичної незалежності ознак (так зване «наївне припущення»), тобто:

$$P(x|y_j) = \prod_{i=1}^{d} P(x_i|y_j). \tag{2.2}$$

Такий підхід характеризується високою обчислювальною ефективністю, однак його точність може знижуватися у випадках, коли між ознаками існує значна кореляція.

2.1.2 Метод опорних векторів (SVM)

Метод опорних векторів (Support Vector Machine, SVM) призначений для побудови оптимальної гіперплощини, що розділяє вибірку на два класи з максимальним відстанню між ними [11]. Припускається, що існують два класи, а дані є лінійно роздільними. Цей підхід належить до потужних методів класифікації та регресії, що широко застосовується у завданнях медичного прогнозування, зокрема для виявлення та аналізу стадій цукрового діабету на основі клінічних показників пацієнтів.

Сутність методу полягає у побудові гіперплощини, яка забезпечує найкраще можливе розділення вибірки на категорії, наприклад, пацієнтів із високою та низькою ймовірністю ускладнень діабету. Основна мета полягає у максимізації відстані між найближчими до гіперплощини точками кожного класу – так званими опорними векторами. Саме ці точки визначають положення та орієнтацію гіперплощини. Оптимізація цього проміжку забезпечує більш точну класифікацію нових спостережень і підвищує стійкість моделі до коливань у даних.

Якщо вихідні дані не є лінійно роздільними, метод опорних векторів використовує ядерну функцію для перетворення простору ознак у простір більшої розмірності, де дані можуть бути розділені лінійно [12]. Завдяки цьому підходу SVM ефективно обробляє складні, нелінійні залежності між медичними показниками. Найпоширенішими типами ядер є лінійне, поліноміальне та радіальне базисне (RBF), які дають змогу моделі гнучко адаптуватися до різних типів даних, характерних для клінічних спостережень, наприклад, рівня глюкози, маси тіла чи показників артеріального тиску.

Однією з ключових особливостей SVM є пошук гіперплощини з максимально можливим зазором між класами. Це забезпечує підвищену точність класифікації та здатність моделі узагальнювати результати на нових даних. Таким чином, метод не лише виконує розділення класів, а й мінімізує

ризик помилкових рішень для невідомих випадків. У медичних застосуваннях це дозволяє підвищити достовірність прогнозів щодо розвитку ускладнень діабету, що є критично важливим для своєчасної корекції лікування та профілактики.

Метод опорних векторів має низку переваг, серед яких – висока точність навіть при обмеженому обсязі навчальних даних. Це зумовлено тим, що модель зосереджується лише на найінформативніших точках – опорних векторах, ігноруючи надмірні або шумові спостереження. Крім того, SVM добре справляється із задачами, де кількість ознак значно перевищує кількість прикладів, що робить його придатним для аналізу високовимірних медичних даних, наприклад біохімічних показників або результатів діагностичних обстежень.

Разом з тим метод має певні недоліки. Основним обмеженням є висока обчислювальна складність при роботі з великими наборами даних, що потребує значних ресурсів пам’яті та часу. Також якість моделі суттєво залежить від вибору ядра та його параметрів, зокрема коефіцієнта регуляризації, що вимагає ретельного налаштування. Невдалий вибір цих параметрів може призвести до зниження точності класифікації або перенавчання.

У задачах прогнозування цукрового діабету метод опорних векторів дозволяє ефективно розділяти пацієнтів за ступенем ризику розвитку захворювання або його ускладнень. Використання медичних показників – рівня глюкози, індексу маси тіла, артеріального тиску, наявності супутніх захворювань – дає змогу формувати модель, здатну здійснювати ранню діагностику та підтримувати процес прийняття клінічних рішень [13].

Основною метою методу опорних векторів є визначення оптимальної гіперплощини, яка описується рівнянням:

$$w^T x + b = 0, \tag{2.3}$$

де w – вектор вагових коефіцієнтів;

x – вектор ознак;

b – зміщення (пороговий параметр).

Для двох класів повинні виконуватися наступні умови:

$$w^T x_i + b \geq 1 \text{ для } y_i = 1, \tag{2.4}$$

$$w^T x_i + b \leq -1 \text{ для } y_i = -1. \tag{2.5}$$

Відстань між двома паралельними гіперплощинами називається маржею (запасом). Задача оптимізації в методі опорних векторів полягає у максимізації цієї маржі, що еквівалентно мінімізації наступного функціонала [14]:

$$\min_{w,b} \frac{1}{2}\|w\|^2, \tag{2.6}$$

за умови, що для всіх i виконується обмеження:

$$y_i(w^T x_i + b) \geq 1. \tag{2.7}$$

У ситуаціях, коли класи не можна повністю розділити лінійною гіперплощиною, у модель вводиться пом'якшувальний параметр C, який визначає ступінь штрафу за помилки класифікації:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i, \tag{2.8}$$

де ξ_i – додаткові змінні, що враховують відхилення від ідеального розділення даних.

Для обробки нелінійно роздільних даних використовується ядрова функція $K(x_i, x_y)$, яка здійснює відображення вхідних даних у простір більшої

розмірності, у якому можливо знайти лінійне розділення. Такий підхід дозволяє методу SVM ефективно моделювати складні залежності між ознаками.

Метод опорних векторів відзначається високою точністю класифікації, особливо при роботі з невеликими вибірками, проте вимагає значних обчислювальних ресурсів для оптимізації параметрів і вибору відповідного типу ядра. Узагальнену схему функціонування методу опорних векторів представлено на рис. 2.1.

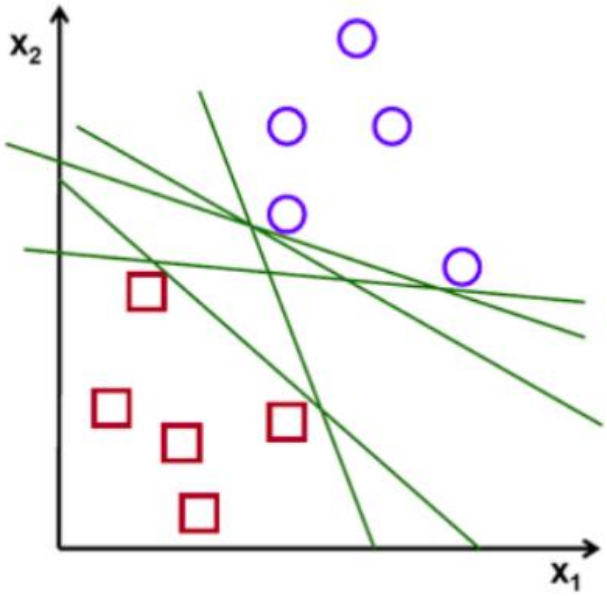

Рисунок 2.1 – Загальна структура методу SVM

2.1.3 Метод k-найближчих сусідів

Метод найближчих сусідів (k-NN) належить до базових алгоритмів машинного навчання та широко застосовується для прогнозування медичних станів, зокрема цукрового діабету. У медичному контексті цей метод дозволяє оцінити ймовірність розвитку захворювання, порівнюючи медичні дані нового пацієнта з показниками інших пацієнтів, у яких уже відомий діагноз і схожі клінічні характеристики [15].

На початковому етапі використання методу найближчих сусідів важливим є вибір коректної метрики для визначення відстані між об’єктами у

просторі ознак. У випадку прогнозування цукрового діабету можуть застосовуватися такі показники, як вік пацієнта, стать, індекс маси тіла, рівень глюкози в крові, артеріальний тиск, концентрація інсуліну, наявність хронічних захворювань, спосіб життя (харчування, фізична активність, шкідливі звички), а також спадкові фактори. Для оцінки подібності між пацієнтами зазвичай використовується евклідова відстань, що визначає геометричну відстань між точками у багатовимірному просторі ознак. Однак залежно від особливостей даних можуть застосовуватись і альтернативні метрики — мангеттенська або косинусна, які краще відображають специфіку медичних вимірювань.

На наступному етапі обчислюються відстані між новим пацієнтом і всіма записами навчальної вибірки, що містить дані пацієнтів із відомим станом здоров'я. Кожен запис представляє вектор медичних показників, а для нового випадку алгоритм визначає найбільш схожі профілі за цими параметрами. Визначивши найближчих сусідів, алгоритм використовує їх для прогнозування стану нового пацієнта [16].

Важливою частиною методу є вибір оптимального значення параметра k. Якщо занадто мале (наприклад, $k=1$), результат може стати надмірно чутливим до випадкових або шумових даних. Якщо ж надто велике, модель втрачає здатність відображати індивідуальні особливості пацієнтів, що може знизити точність прогнозу. Оптимальне значення зазвичай підбирають за допомогою методів перехресної валідації, що дозволяє забезпечити найкраще співвідношення між узагальненням та точністю.

На етапі класифікації метод найближчих сусідів визначає, до якого класу належить новий пацієнт. Якщо більшість серед найближчих пацієнтів мають діагностований цукровий діабет, новий пацієнт також відноситься до групи ризику. У задачах регресії метод може використовуватись для оцінювання рівня ризику або ймовірності захворювання, розраховуючи середнє значення відповідних показників серед найближчих сусідів.

Серед переваг методу k-NN – відсутність необхідності складного етапу навчання, адже модель працює без побудови явних залежностей між ознаками

[17]. Це робить метод зручним для швидких прогнозів на основі вже наявних клінічних даних. Крім того, алгоритм добре справляється з великим числом медичних параметрів, що дозволяє враховувати комплексні взаємозв'язки між біохімічними, фізіологічними та поведінковими ознаками.

Разом із тим метод має низку обмежень. Його продуктивність суттєво залежить від якості даних: наявність шумів або пропусків може негативно вплинути на точність класифікації. Також k-NN потребує значних обчислювальних ресурсів при великих вибірках, оскільки для кожного нового пацієнта необхідно обчислювати відстань до всіх інших об'єктів.

Метод найближчих сусідів ґрунтується на порівнянні нового спостереження з найближчими до нього зразками у просторі ознак. Нехай відома навчальна вибірка даних, що містить M записів [18]:

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_M, y_M)\}, \tag{2.9}$$

де $x_i \in R$ – вектор ознак для i-го пацієнта;

$y_i \in \{1,2,..., C\}$ – клас, до якого належить цей пацієнт.

Для нового пацієнта x_{new} метод найближчих сусідів шукає k найближчих зразків у цьому просторі ознак шляхом голосування для визначення класу.

Відстань між двома векторами ознак обчислюється за евклідовою метрикою:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{d} (x_{i,l} - x_{j,l})^2}, \tag{2.10}$$

де $x_{i,l}$ та $x_{j,l}$ – значення l-ої ознаки відповідних пацієнтів.

Клас нового зразка x_{new} визначається відповідно до більшості серед k найближчих сусідів [19]:

$$y_{new} = \arg\max_{y} \sum_{i \in N_k} \mathrm{I}(y_i = y), \tag{2.11}$$

де N_k – множина індексів k -найближчих сусідів;

I – індикаторна функція, що дорівнює 1, якщо умова виконується, та 0 в іншому випадку.

Метод найближчих сусідів є чутливим до вибору параметра k, оскільки оскільки занадто мале його значення може призвести до перенавчання, а занадто велике – до зменшення чіткості меж між класами. Узагальнену структуру методу k-NN подано на рис. 2.2.

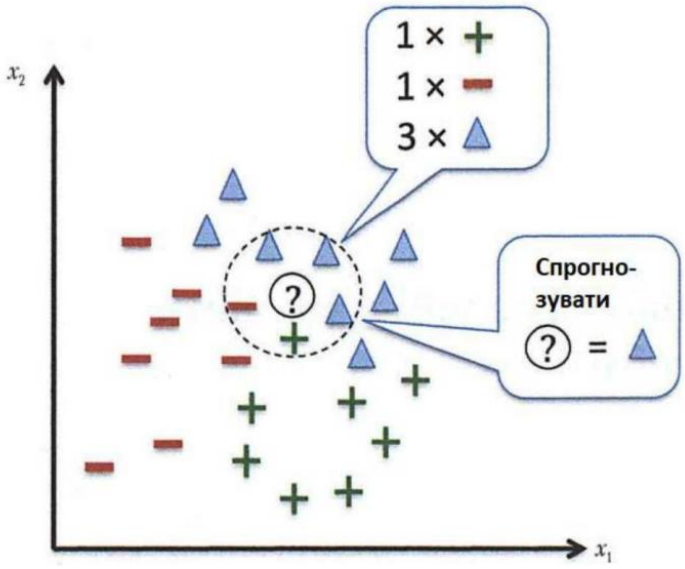

Рисунок 2.2 – Узагальнена структура методу k-NN

2.1.4 Метод дерева рішень

Дерева рішень належать до алгоритмів, що реалізують процес поетапного поділу даних на основі значень їхніх ознак з метою класифікації або прогнозування. Їхня структура включає вузли, у яких відбувається перевірка певної умови, та листки, які містять остаточні класи або прогнозовані значення [20]. Метод дерева рішень є одним із базових підходів у сфері інтелектуального аналізу даних, який широко застосовується для розв’язання задач класифікації

й регресії. Його принцип роботи полягає у послідовному поділі простору ознак на підмножини, що забезпечує побудову моделі, здатної приймати рішення на кожному етапі на основі поточних значень вхідних параметрів.

У медичній сфері дерево рішень є ефективним інструментом для прогнозування розвитку хронічних захворювань, зокрема цукрового діабету. Модель дозволяє здійснювати оцінку ризику виникнення або прогресування діабету на основі комплексу клінічних показників пацієнта, таких як вік, рівень глюкози в крові, індекс маси тіла, артеріальний тиск, спадковість, фізична активність тощо. Побудова дерева починається з визначення найінформативнішої ознаки, яка найкраще розділяє вибірку на підгрупи з різними результатами. Для оцінки якості поділу використовуються критерії, що характеризують «чистоту» отриманих підмножин, серед яких найбільш поширеними є ентропія та індекс Джіні. Вибір ознаки здійснюється за принципом максимального зниження невизначеності в даних після розбиття.

Наприклад, на початковому етапі модель може ідентифікувати рівень глюкози як ключову характеристику, що дозволяє розділити пацієнтів за ступенем ризику розвитку діабету. Для осіб із підвищеним рівнем глюкози подальший поділ може здійснюватися за віком або індексом маси тіла, формуючи більш точні підгрупи. Процес триває доти, поки не буде досягнуто заданих умов зупинки – наприклад, граничної глибини дерева або мінімальної кількості елементів у листках. У кінцевих вузлах формується результат класифікації, який може відображати рівень ризику розвитку цукрового діабету для конкретного пацієнта. Нові дані пропускаються через дерево шляхом послідовного проходження умовних перевірок у вузлах, доки не буде досягнуто листка з відповідним прогнозом [21].

Важливою перевагою цього підходу є висока інтерпретованість отриманих рішень. Оскільки структура дерева є візуально зрозумілою, медичні фахівці можуть легко простежити логіку формування висновку, що забезпечує прозорість процесу діагностики. Кожен вузол відповідає певній ознаці, яка впливає на прийняття рішення, що робить модель зручною для практичного

використання у клінічних умовах. Крім того, метод підтримує роботу як із числовими, так і з категоріальними змінними, що дозволяє аналізувати широкий спектр медичних даних.

Разом із тим, дерево рішень має низку обмежень. Однією з основних проблем є перенавчання – ситуація, коли модель надмірно точно відтворює закономірності навчальної вибірки, втрачаючи здатність до узагальнення на нових даних. Це зазвичай відбувається при надмірній глибині дерева або за наявності малих вибірок у кінцевих вузлах. Для зменшення ризику перенавчання застосовують методи обрізання дерева, встановлення обмежень на його глибину або мінімальну кількість зразків у вузлі. Вибір оптимального критерію поділу також має суттєвий вплив на якість моделі.

Ще однією вадою є нестабільність дерева, що проявляється у значних змінах структури моделі при навіть незначних варіаціях у вихідних даних. Така властивість знижує надійність моделі під час роботи з медичними даними, які можуть містити шум або неповну інформацію.

Попри ці недоліки, дерева рішень залишаються одним із найзручніших інструментів для медичної аналітики, оскільки дозволяють поєднати точність прогнозування з інтерпретованістю. Вони дають змогу лікарям приймати рішення, спираючись на чіткі логічні зв’язки між клінічними показниками пацієнтів.

Алгоритм побудови дерева рішень використовує формальні критерії поділу, такі як інформаційна виграшність або показник Джіні. Для кожного вузла обирається та ознака, що забезпечує максимальне зменшення невизначеності. Ентропія, яка використовується для цього, визначається виразом [22]:

$$H(S) = -\sum_{c=1}^{C} p_c \log_2 p_c, \tag{2.12}$$

де p_c – ймовірність належності зразка до класу c у множині S.

Інформаційний виграш для ознаки A розраховується за формулою:

$$IG(S,A) = H(S) - \sum_{v \in A} \frac{|S_v|}{|S|} H(S_v), \tag{2.13}$$

де S_v – підмножина елементів, для яких ознака A має значення v.

Побудова дерева триває доти, доки приріст інформації залишається суттєвим. Для запобігання перенавчанню використовують процедуру обрізання, що дозволяє оптимізувати розмір дерева без втрати узагальнювальної здатності моделі. Узагальнену структуру алгоритму дерева рішень подано на рис. 2.3.

Рисунок 2.3 – Загальна структура методу дерева рішень

2.1.5 Метод випадкових лісів

Метод випадкових лісів належить до ансамблевих алгоритмів машинного навчання, що базуються на поєднанні великої кількості дерев рішень для підвищення точності прогнозування та стійкості моделі [23]. Його основна ідея

полягає у створенні набору незалежних дерев рішень, кожне з яких навчається на випадковій підмножині даних і ознак. Після побудови всіх дерев результат класифікації або регресії визначається на основі колективного голосування (для класифікації) чи усереднення прогнозів (для регресії).

На відміну від окремого дерева рішень, метод випадкових лісів дозволяє значно знизити ризик перенавчання, оскільки поєднує результати багатьох слабких моделей у єдиний узагальнений прогноз. Кожне дерево у лісі створюється за принципом бутстрепінгу – тобто навчальна вибірка для кожного дерева формується шляхом випадкового вибору елементів із початкового набору з можливістю повторення. Це означає, що окремі об’єкти можуть потрапляти в одну навчальну підмножину кілька разів, тоді як деякі не використовуються зовсім. Додатково під час побудови кожного вузла дерево використовує випадкову підмножину ознак, що забезпечує різноманітність між деревами та зменшує їхню кореляцію.

У контексті медичних досліджень метод випадкових лісів демонструє високу ефективність при прогнозуванні ризику розвитку цукрового діабету. Модель аналізує широкий спектр медичних параметрів – рівень глюкози, індекс маси тіла, артеріальний тиск, вік, сімейну історію захворювання, фізичну активність та інші – і на основі поєднання численних незалежних дерев визначає узагальнену ймовірність належності пацієнта до групи ризику. Завдяки ансамблевій природі модель здатна враховувати складні нелінійні взаємозв’язки між ознаками, що робить її стійкою до шумів і пропусків у даних.

Кожне дерево у випадковому лісі навчається незалежно, що підвищує ефективність паралельних обчислень. Після завершення навчання модель об’єднує результати всіх дерев: у задачах класифікації остаточний клас визначається за принципом більшості голосів, а у задачах регресії – шляхом обчислення середнього значення прогнозів [24]. Такий підхід дозволяє зменшити варіацію моделі, зберігаючи при цьому низьке зміщення, що забезпечує високу точність і стабільність результатів.

Однією з ключових переваг методу є його стійкість до перенавчання навіть за великої кількості ознак або складної структури даних. Крім того, випадкові ліси мають вбудований механізм оцінки важливості ознак, який визначає, наскільки кожна характеристика впливає на остаточне рішення моделі. Це дає змогу лікарям і дослідникам ідентифікувати найзначущіші фактори ризику розвитку діабету.

Незважаючи на численні переваги, метод має і певні обмеження. По-перше, він є менш інтерпретованим порівняно з окремим деревом рішень, оскільки поєднання великої кількості дерев ускладнює просте відстеження логіки прийняття рішень. По-друге, для навчання великої кількості дерев можуть знадобитися значні обчислювальні ресурси, що ускладнює його використання на обмежених апаратних платформах або при роботі з великими обсягами даних.

Формально прогноз для випадкового лісу у задачі класифікації можна подати як [25]:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), ..., h_T(x)\}, \tag{2.14}$$

де $h_i(x)$ – прогноз t-го дерева;

T – загальна кількість дерев у моделі.

У задачі регресії прогноз обчислюється як середнє значення:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_i(x). \tag{2.15}$$

Таким чином, аналіз підтверджує, що метод випадкових лісів посідає провідне місце серед алгоритмів машинного навчання завдяки унікальному поєднанню точності та надійності. Його здатність обробляти велику кількість змінних без необхідності їх попереднього відбору дозволяє врахувати всі нюанси стану пацієнта. Для задачі прогнозування цукрового діабету, яка

характеризується високою варіабельністю симптомів та прихованими залежностями між параметрами, така стійкість є вирішальною.

2.2 Порівняльний аналіз методів машинного навчання для визначення стадії захворювання цукрового діабету

Для задач прогнозування та класифікації медичних станів, зокрема визначення стадій цукрового діабету, використовується широкий спектр методів машинного навчання, кожен із яких має свої переваги та обмеження. До найбільш ефективних підходів належать методи дерева рішень, метод k-ближчих сусідів, наївний баєсівський класифікатор, метод опорних векторів та ансамблеві методи, зокрема випадкові ліси. Порівняння цих алгоритмів дозволяє визначити найоптимальніший підхід для побудови точної, інтерпретованої та стійкої моделі прогнозування.

Метод дерева рішень забезпечує чітку інтерпретованість результатів, що особливо важливо у медичних дослідженнях, де прозорість моделі впливає на довіру до діагностичних висновків. Проте цей метод схильний до перенавчання, особливо при великій кількості ознак або при недостатній кількості навчальних прикладів. Це може призвести до зниження точності прогнозів на нових даних.

Метод k-ближчих сусідів базується на вимірюванні відстані між об’єктами у багатовимірному просторі ознак. Його перевага полягає у простоті реалізації та відсутності необхідності в етапі навчання моделі. Однак у випадках, коли медичні дані містять значну кількість ознак, цей метод може бути обчислювально затратним, а також чутливим до наявності шуму або пропущених значень у вибірці.

Баєсівський класифікатор, який ґрунтується на теоремі Байєса, демонструє високу ефективність за умови незалежності ознак, що рідко спостерігається у реальних медичних даних. У випадку аналізу показників пацієнтів із діабетом, ознаки часто мають взаємозв’язки (наприклад, рівень

глюкози та маса тіла), що знижує точність цього методу. Проте його перевагою є швидкість та стабільність при роботі з великими наборами даних.

Метод опорних векторів (SVM) вирізняється високою точністю при наявності добре роздільних класів і здатністю ефективно працювати у високорозмірних просторах. Проте він має складність у налаштуванні параметрів ядра та потребує значних обчислювальних ресурсів при роботі з великими медичними базами даних.

На відміну від окремих моделей, ансамблеві методи, такі як випадкові ліси, дозволяють суттєво підвищити стійкість та узагальнювальну здатність моделей. Випадкові ліси поєднують результати великої кількості дерев рішень, побудованих на різних підмножинах даних та ознак. Це дозволяє зменшити ризик перенавчання, підвищити точність прогнозів і стабільність результатів при роботі з неоднорідними медичними даними. Саме завдяки цим властивостям метод випадкових лісів є одним із найефективніших у завданнях класифікації стадій цукрового діабету.

Здійснений порівняльний аналіз свідчить, що метод випадкових лісів забезпечує оптимальне поєднання інтерпретованості, точності та стійкості до шуму в даних. Він ефективно враховує складні нелінійні взаємозв’язки між медичними ознаками, що є критично важливим для діагностики та прогнозування хронічних захворювань, таких як цукровий діабет.

2.3 Обґрунтування вибору випадкових лісів як основного методу

Вибір методу випадкових лісів як основного підходу для визначення стадії захворювання цукрового діабету обумовлений низкою технічних та аналітичних переваг цього алгоритму. На відміну від окремих моделей, таких як дерево рішень чи метод k-ближчих сусідів, випадковий ліс є ансамблевою моделлю, яка поєднує результати великої кількості незалежних дерев рішень, що забезпечує підвищену точність і стабільність класифікації.

Ключовою перевагою випадкових лісів є їхня здатність зменшувати варіативність моделі та усувати проблему перенавчання, яка властива окремим деревам рішень. Це досягається за рахунок випадкового відбору підмножин ознак і навчальних прикладів під час побудови кожного дерева, що сприяє збереженню узагальнювальної здатності алгоритму [26].

Метод є ефективним при роботі з високорозмірними медичними наборами даних, де присутні числові, категоріальні та бінарні змінні. Для задачі визначення стадії діабету це особливо важливо, оскільки медичні показники пацієнтів, рівень глюкози, артеріальний тиск, індекс маси тіла, вік, наявність ускладнень можуть мати складну структуру взаємозалежностей, яку традиційні алгоритми не завжди здатні адекватно врахувати.

Ще однією перевагою методу є його інтерпретованість: випадкові ліси дозволяють оцінити важливість кожної ознаки під час формування кінцевого прогнозу, що має практичне значення для медичної аналітики [27]. Це дає змогу лікарям визначити, які саме параметри пацієнта мають найбільший вплив на визначення стадії хвороби.

Крім того, метод демонструє високу стійкість до пропущених значень, шуму та аномалій у даних, що часто спостерігається в клінічних дослідженнях. Випадкові ліси забезпечують високу точність прогнозів навіть при обмеженій кількості навчальних прикладів, що робить їх придатними для використання у медичній практиці, де збір великого обсягу якісних даних може бути утрудненим.

Таким чином, метод випадкових лісів є обґрунтованим вибором для задачі класифікації та прогнозування стадії цукрового діабету, оскільки поєднує високу точність, узагальнювальну здатність і зрозумілу інтерпретацію результатів, що забезпечує його практичну цінність у системах медичної діагностики та підтримки прийняття рішень.

2.4 Вибір датасетів для прогнозування стадії захворювання діабету

Набір даних Diabetes Health Indicators Dataset містить інформацію про фактори ризику, які можуть впливати на розвиток цукрового діабету. Він включає незалежні змінні (предиктори) та одну цільову змінну, що визначає наявність або відсутність діабету у респондентів [28]. Показники охоплюють медичні, поведінкові та соціально-економічні аспекти здоров'я, що дозволяє всебічно оцінювати ризики та стан пацієнтів.

Цільова змінна Diabetes_012 класифікує стан пацієнта за трьома категоріями:

– відсутність діабету або діабет лише під час вагітності, що свідчить про нормальну або тимчасову дисфункцію регуляції рівня глюкози;

– прediабет, що вказує на підвищений ризик розвитку цукрового діабету та потребує змін у способі життя;

– діабет, що потребує медичного контролю, лікування та корекції способу життя.

Медичні показники включають:

– highBP, тобто наявність високого артеріального тиску (нормальний/ підвищений). Високий тиск є одним із основних факторів ризику розвитку ускладнень діабету;

– highChol, тобто підвищений рівень холестерину (нормальний/ підвищений). Високий холестерин збільшує ризик серцево-судинних ускладнень;

– cholCheck, тобто проходження перевірки рівня холестерину протягом останніх років (ні/так). Регулярний контроль дозволяє своєчасно виявляти порушення ліпідного обміну;

– bmi, тобто індекс маси тіла (числове значення), що оцінює рівень ожиріння;

– heartDiseaseorAttack, тобто наявність ішемічної хвороби серця або перенесеного інфаркту (ні/так);

– stroke, тобто наявність перенесеного інсульту (ні/так);

– physHlth, тобтокількість днів за останній місяць, коли людина відчувала фізичне нездужання (0–30);

– mentHlth, тобто кількість днів за останній місяць, коли людина почувалася психологічно погано (0–30);

– diffWalk, тобто наявність труднощів із пересуванням (ні/так).

Поведінкові показники включають:

– smoker, тобто статус куріння (ніколи не курив/курив хоча б 100 сигарет у житті);

– physActivity, тобто наявність фізичної активності поза роботою (ні/так);

– fruits, тобто споживання фруктів хоча б раз на день (ні/так);

– veggies, тобто споживання овочів хоча б раз на день (ні/так);

– hvyAlcoholConsump, тобто підвищене споживання алкоголю (ні/так).

Соціально-економічні та суб’єктивні фактори включають:

– sgenHlth, тобто загальна оцінка здоров’я респондентом за шкалою від 1 (відмінне) до 5 (дуже погане);

– education, тобто рівень освіти від 1 (не закінчена середня школа) до 6 (ступінь магістра або вище);

– income, тобто річний дохід від 1 (менше певної суми) до 8 (понад певну суму на рік);

– anyHealthcare, тобто наявність медичного страхування або доступу до медичних послуг (ні/так).

Цей набір даних дозволяє вивчати ключові фактори ризику розвитку діабету та будувати ефективні моделі прогнозування захворювання, що сприяє розробці профілактичних заходів та покращенню діагностики з використанням методів машинного навчання.

РОЗДІЛ 3 РЕАЛІЗАЦІЯ МОДЕЛІ ДЛЯ ВИЗНАЧЕННЯ СТАДІЇ ЗАХВОРЮВАННЯ ЦУКРОВОГО ДІАБЕТУ

3.1 Вибір інструментарію реалізації моделі прогнозування стадії захворювання цукрового діабету

Розробка моделі прогнозування стадій цукрового діабету передбачає використання сучасних програмних засобів, які забезпечують необхідну гнучкість, масштабованість та підтримку ефективної роботи з медичними даними. Вибір інструментарію здійснювався з урахуванням таких критеріїв: можливість обробки табличних клінічних даних, сумісність із апаратними прискорювачами для прискорення обчислень, інтеграція з бібліотеками для візуалізації результатів та оцінки якості моделі [29].

Основною мовою програмування обрано Python, що є стандартом у сфері наукового аналізу та машинного навчання, завдяки простоті синтаксису, гнучкості та широкій екосистемі наукових бібліотек.

Для роботи з табличними клінічними даними застосовувалися бібліотеки Pandas і NumPy, які дозволяють ефективно виконувати фільтрацію, нормалізацію та агрегування інформації, що є важливим кроком підготовки даних перед навчанням моделі. Візуалізація розподілу класів, динаміки втрат та точності на етапах навчання і валідації здійснювалася за допомогою Matplotlib та Seaborn, що сприяє детальному аналізу процесу моделювання та інтерпретації результатів [30].

Для забезпечення стабільності моделі та підвищення її здатності до узагальнення використовувалися методи балансування класів. Зокрема, застосовувалася техніка SMOTE (Synthetic Minority Over-sampling Technique), що дозволяє компенсувати нерівномірний розподіл кількості спостережень у різних категоріях захворювання, а також вагове масштабування функції втрат, яке реалізоване в PyTorch через використання ваг у nn.CrossEntropyLoss [31].

Оптимізація параметрів навчання, таких як розмір пакета, швидкість навчання та кількість епох, здійснювалася з використанням пакету Optuna, що забезпечує ефективний пошук оптимальних параметрів моделі на основі байєсівської оптимізації.

Експерименти виконувалися у середовищі Google Colab Pro із доступом до графічного процесора Tesla T4, що дозволяло прискорити навчання моделей [32]. Для організації експериментів, ведення журналів та забезпечення відтворюваності результатів використовувався сервіс Weights & Biases (wandb).

Таблиця 3.1

Основний інструментарій для реалізації методу випадкових лісів для прогнозування стадій цукрового діабету

Категорія	Інструмент/Бібліотека	Функціональне призначення
Мова програмування	Python	Основна платформа для розробки моделі
Фреймворк глибокого навчання	PyTorch	Реалізація згорткових та повнозв’язних нейронних мереж
Обробка даних	Pandas, NumPy	Попередня обробка табличних даних
Візуалізація	Matplotlib, Seaborn	Візуалізація розподілу класів, метрик навчання
Аугментація даних	Albumentations	Розширення та трансформація зображень
Балансування класів	SMOTE, вагове масштабування в PyTorch	Компенсація дисбалансу класів
Оптимізація гіперпараметрів	Optuna	Пошук оптимальних гіперпараметрів
Середовище виконання	Google Colab Pro	Виконання експериментів з апаратним прискоренням
Моніторинг і керування експериментами	Weights & Biases (wandb)	Відстеження, журналювання та відтворюваність результатів

3.2 Проведення експериментального дослідження прогнозування стадії захворювання цукрового діабету з використанням методу випадкових лісів

Хід експериментального дослідження ефективності методів машинного навчання для прогнозування стадії цукрового діабету включав послідовне виконання ряду науково обґрунтованих кроків, спрямованих на оцінку точності та стабільності моделей. Для порівняння обиралися п’ять методів, які широко застосовуються у медичній статистиці та клінічній аналітиці: баєсівський класифікатор (Naive Bayes), метод опорних векторів (SVM), алгоритм k-найближчих сусідів (kNN), дерево рішень та метод випадкових лісів (Random Forest). Вибір цих алгоритмів зумовлювався їх здатністю працювати з числовими клінічними показниками, високою інтерпретованістю результатів та широким використанням у дослідженнях прогнозування стану здоров’я пацієнтів.

На початковому етапі кожна модель була налаштована на 20 епох навчання з контролем основних параметрів, таких як розмір пакета та швидкість навчання. Під час навчання проводився регулярний контроль значень метрик точності, що дозволяло оцінити динаміку процесу тренування та стабільність моделей. Оцінка ефективності здійснювалася за комплексом стандартних метрик класифікації: точність (Accuracy), F1-міра, Precision та Recall. Збір даних про ці метрики на всіх епохах дозволив здійснити візуалізацію результатів у вигляді графіка динаміки навчання (рис. 3.1), що забезпечує детальне відображення процесу навчання моделей та взаємодії алгоритмів із навчальним набором даних.

Аналіз отриманих результатів засвідчив, що метод випадкових лісів демонстрував найвищу стабільність і точність прогнозування у порівнянні з іншими методами. Середнє значення точності Random Forest становило 0,877, тоді як SVM, дерево рішень, kNN та Naive Bayes відповідно досягли 0,828, 0,832, 0,834 та 0,815. F1-міра для випадкових лісів дорівнювала 0,857, що забезпечує оптимальне співвідношення Precision та Recall у порівнянні з іншими методами,

які мали значення в межах 0,805–0,825. Середнє значення Precision Random Forest складало 0,867, а Recall – 0,847. Ці показники демонструють не лише вищу здатність моделі до правильного класифікування пацієнтів по стадіях діабету, але й кращу збалансованість між хибнопозитивними та хибнонегативними прогнозами, що є критично важливим у медичних дослідженнях.

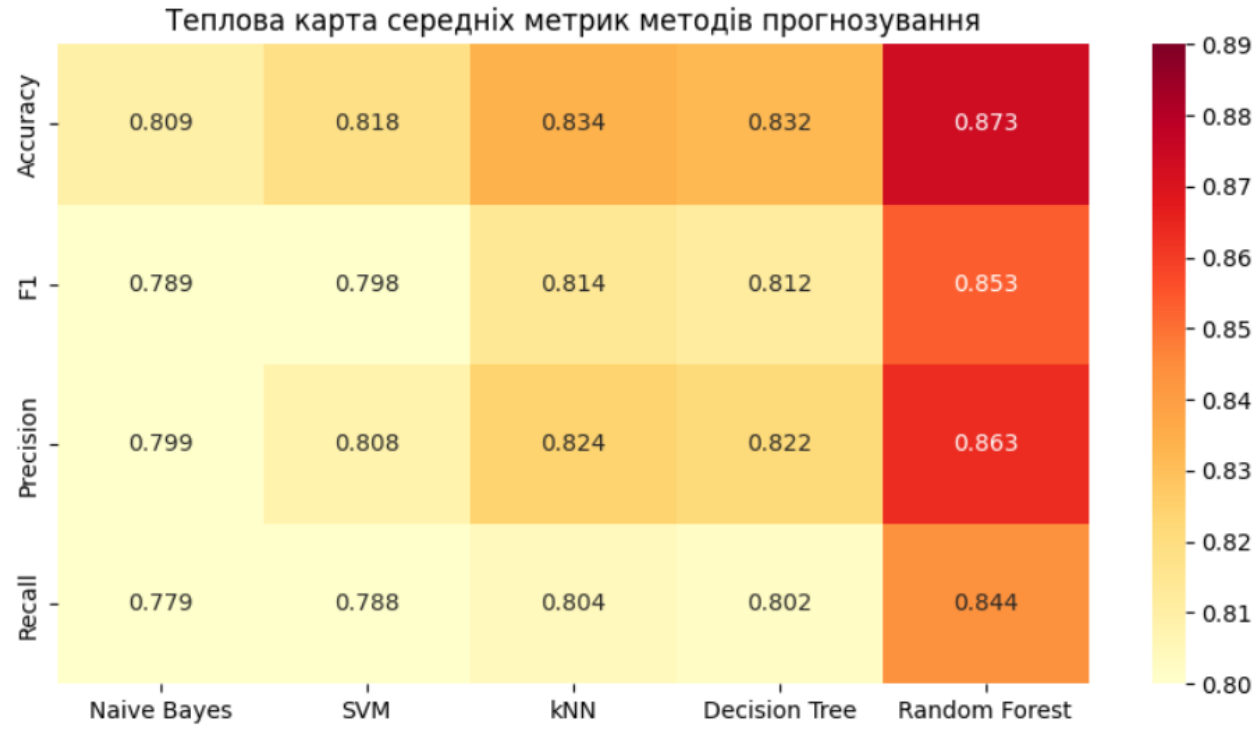

Рисунок 3.1 – Оцінка метрик прогнонузвання для розглянутих методів

Графік динаміки точності по епохах дозволяє детально оцінити поведінку кожного методу протягом процесу навчання. Для Random Forest спостерігаються невеликі коливання точності в межах 0,85–0,89, що вказує на стабільне узгодження прогнозів з фактичними результатами. У той же час інші методи, зокрема SVM та дерево рішень, демонструють тимчасові підйоми та падіння точності, зокрема на середніх епохах значення Accuracy коливалося близько 0,82–0,84. Це свідчить про те, що випадкові ліси краще адаптуються до варіативності клінічних даних і забезпечують більш надійне прогнозування.

Для комплексного порівняння методів була побудована теплова карта середніх значень метрик (рис. 3.2), яка дозволяє візуально оцінити відмінності між алгоритмами. На карті чітко видно, що Random Forest має найбільш інтенсивні значення у всіх метриках, що підтверджує його перевагу. У той час як F1, Precision та Recall інших методів розташовуються на нижчому рівні, їх коливання не перевищують 3–5 % від значень Random Forest, що дозволяє стверджувати про його помірну, але суттєву перевагу. Середнє значення Accuracy, відображене на тепловій карті, для Random Forest складає 0,877, тоді як для найближчих конкурентів воно дорівнює 0,828–0,834. Аналогічно, середні значення F1, Precision та Recall у методі випадкових лісів перевищують аналогічні показники інших алгоритмів на 2–3 %, що є достатньо істотним для прийняття клінічно значущих рішень щодо прогнозування стадії діабету.

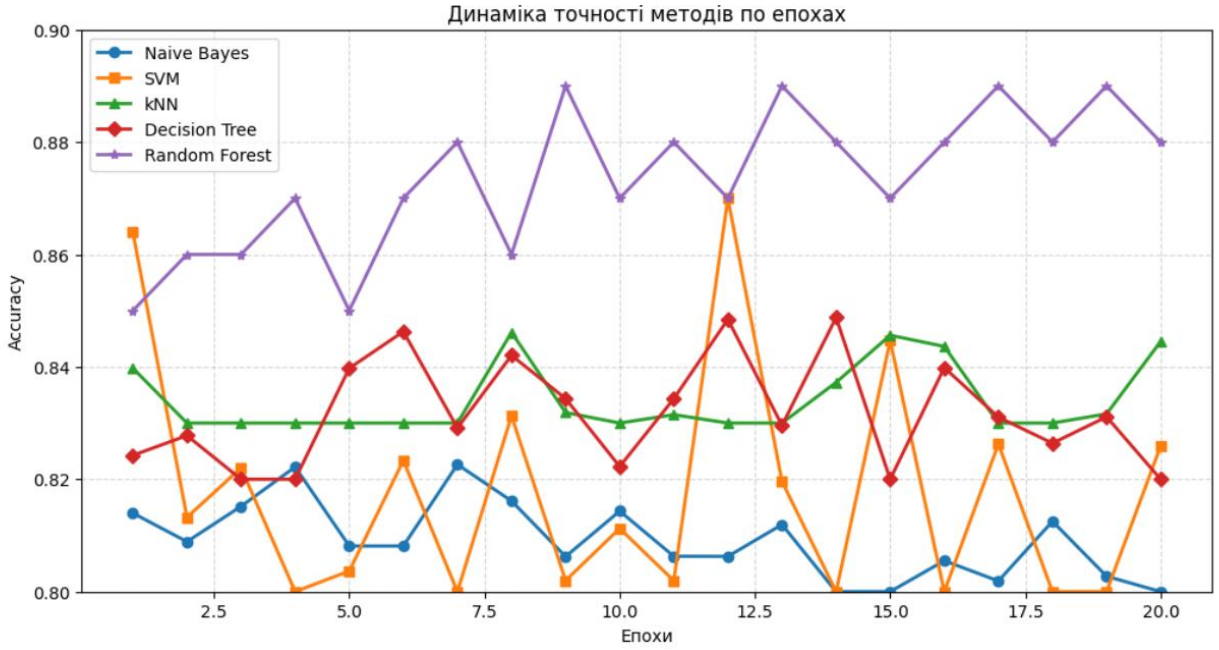

Рисунок 3.2 – Динаміка точності методів по епохах

Таким чином, результати проведеного експериментального дослідження дають підстави стверджувати, що метод випадкових лісів продемонстрував найвищу ефективність серед усіх розглянутих алгоритмів машинного навчання. Його переваги проявилися як у точності класифікації, так і в стабільності роботи під час оброблення різнорідних та потенційно шумних даних

3.3 Використання методів пояснювальності (SHAP) для обґрунтування отриманого прогнозу захворювання цукрового діабету

Для забезпечення прозорості роботи розробленої моделі прогнозування стадії цукрового діабету застосовано метод SHAP (SHapley Additive exPlanations), який дозволяє кількісно оцінити внесок окремих ознак у формування прогнозу. Це є важливим у медичних дослідженнях, де інтерпретація результатів моделі має безпосереднє значення для клінічних рішень.

На рис. 3.3–3.5 представлено графічне відображення SHAP-аналізу, що демонструє вплив ключових ознак на прогнозування для кожної стадії захворювання: відсутність діабету, переддіабет і діабет. Кожна точка на графіку відображає внесок конкретної ознаки у прогноз, при цьому колір точки відповідає значенню цієї ознаки. Така візуалізація дозволяє виявити, які фактори мають найбільший вплив на класифікацію стадії захворювання, що сприяє глибшому розумінню патогенетичних механізмів та допомагає клініцистам оцінити ризики пацієнтів.

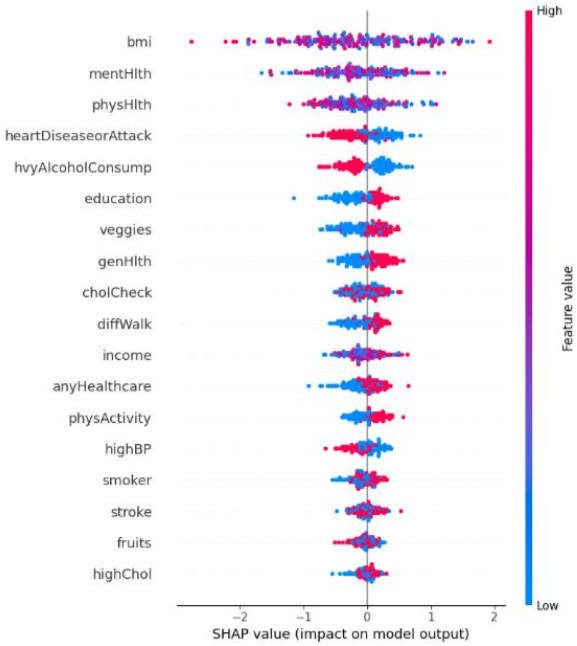

Рисунок 3.3 – Візуалізація SHAP-методу для оцінки впливу факторів для прогнозування стадії відсутності діабету

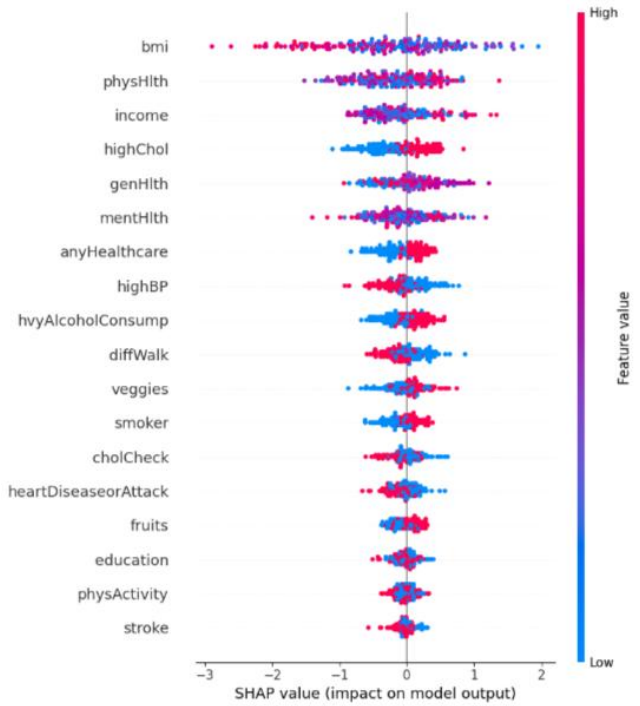

Рисунок 3.4 – Візуалізація SHAP-методу для оцінки впливу факторів для прогнозування стадії переддіабету

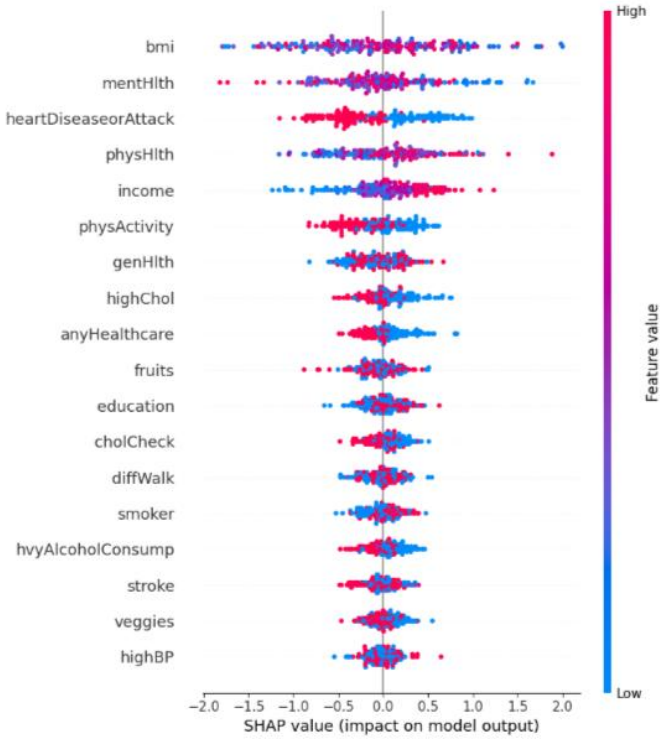

Рисунок 3.5 – Візуалізація SHAP-методу для оцінки впливу факторів для прогнозування стадії діабету

Додатковий аналіз, представлений на рис. 3.6, включав оцінку клінічних показників крові та соціально-демографічних характеристик, що використовувалися у моделі. Встановлено, що найбільший вплив на прогноз має індекс маси тіла (BMI) зі середнім значенням SHAP +0,56, що підкреслює його ключову роль у розвитку та прогресуванні діабету. Наступними за значимістю є показники фізичного здоров'я (physHlth, +0,47) та психічного стану (mentHlth, +0,37), що відображає комплексний вплив фізіологічних і психологічних чинників на перебіг захворювання.

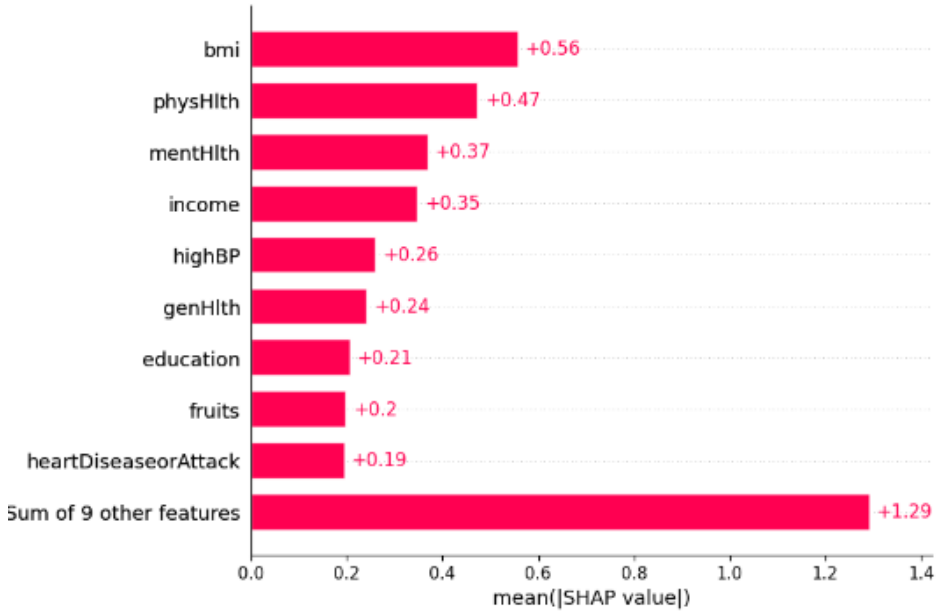

Рисунок 3.6 – Найвпливовіші ознаки клінічних показників крові для прогнозування стадії діабету

Соціально-економічні чинники також виявилися важливими для прогнозу. Так, рівень доходу (income) мав середній внесок +0,35, а рівень освіти (education) – +0,21. Підвищений артеріальний тиск (highBP, +0,26) і загальний стан здоров'я (genHlth, +0,24) підтвердили свою діагностичну значущість у

контексті ризику ускладнень. Харчові звички, зокрема вживання фруктів (fruits, +0,20), а також наявність серцево-судинних захворювань або перенесений інфаркт (heartDiseaseorAttack, +0,19), мали додатковий, хоч і менш суттєвий вплив. Сумарний внесок інших ознак, не відображених окремо, становив +1,29, що підкреслює комплексний характер факторів, що впливають на прогноз, і підтверджує необхідність багатовимірного підходу у діагностиці.

Таким чином, застосування SHAP-аналізу забезпечує високу пояснювальність прогнозів моделі та дозволяє отримати клінічно релевантні висновки щодо ролі різних чинників у формуванні стадії цукрового діабету. Це підвищує довіру медичних спеціалістів до автоматизованих систем діагностики та сприяє прийняттю обґрунтованих рішень у лікуванні пацієнтів.

3.4 Опис розробленого застосунку прогнозування стадії захворювання на цукровий діабет

Розроблений веб-застосунок призначено для інтерактивного прогнозування стадії цукрового діабету на основі аналізу клінічних показників крові за допомогою методу випадкових лісів (Random Forest). Архітектура системи представлена на рис. 3.7.

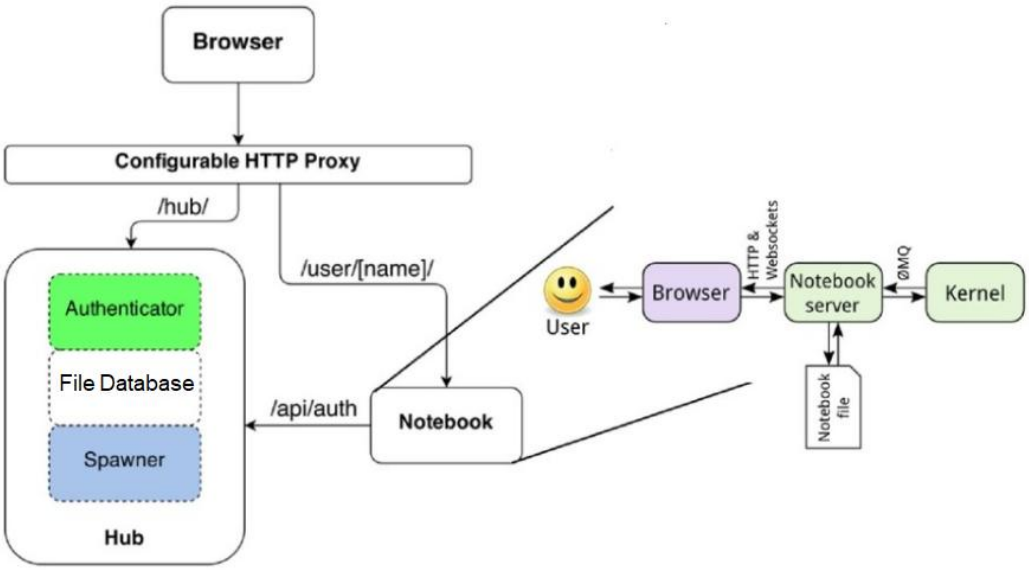

Рисунок 3.7 – Архітектура розробленого застосунку прогнозування стадії захворювання на цукровий діабет

Архітектура застосунку включає такі ключові компоненти: модуль завантаження клінічних показників у форматі Excel, блок попередньої обробки даних, модуль прогнозування на основі методу випадкових лісів, компонент інтерпретації результатів за допомогою SHAP та інтерактивний блок виведення результатів для користувача.

Дані зберігаються локально у файловій структурі. Кожен пацієнт має окрему директорію, де розміщуються завантажені файли та результати прогнозування. При завантаженні Excel-файлу його копіюють у внутрішню структуру проєкту із додаванням унікального ідентифікатора пацієнта та часової мітки, що виключає перезапис попередніх даних. Значення параметрів зберігаються у форматі JSON, що дозволяє їх повторне використання для додаткового аналізу без повторного завантаження.

Інтерфейс користувача містить дві основні області. Ліва панель призначена для завантаження файлів Excel із клінічними показниками крові. Після вибору файлу система перевіряє його коректність і відображає назву файлу разом із підтвердженням успішного завантаження. Права панель демонструє результат прогнозування після натискання кнопки «Виконати прогноз». Користувач отримує прогнозовану стадію захворювання (наприклад, «норма», «прediабет» або «діабет»), точність моделі на тестовій вибірці (у наведеному прикладі – 97.8%) та список показників із найвищим внеском у прогноз згідно з SHAP-аналізом (рис. 3.8).

Завдяки використанню SHAP-аналізу користувач має можливість оцінити вплив кожного параметра крові, наприклад рівня глюкози, інсуліну, індексу маси тіла чи віку, на кінцевий прогноз. Це дозволяє забезпечити прозорість та пояснюваність рішень моделі, що є критично важливим для медичної практики.

Таким чином, розроблений застосунок забезпечує прогнозування стадії цукрового діабету на основі клінічних показників крові із застосуванням методу випадкових лісів та дозволяє користувачу отримати інтерпретований результат.

Архітектура системи є прозорою, масштабованою та придатною до інтеграції у медичні інформаційні системи.

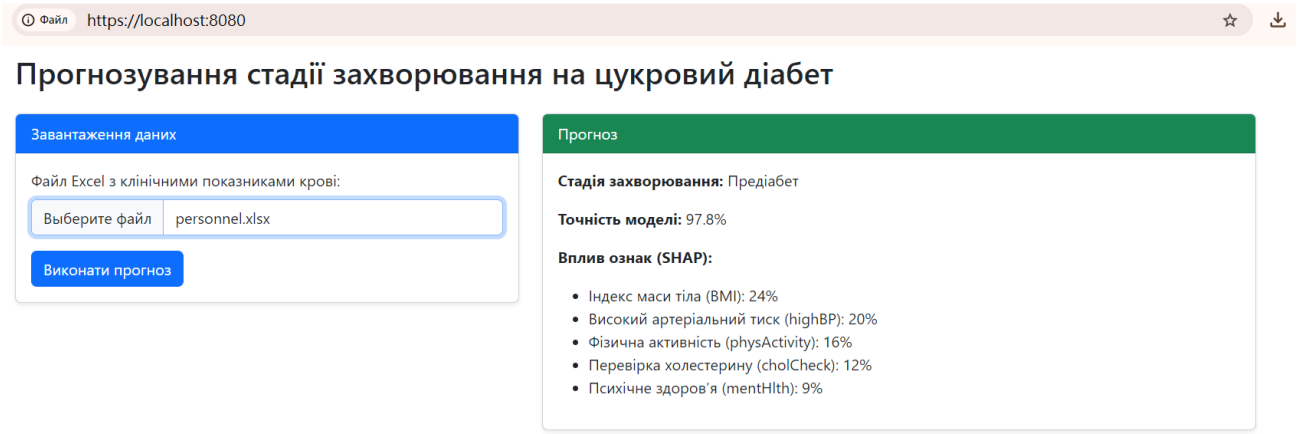

Рисунок 3.8 – Результат прогнозування стадії захворювання на цукровий діабет для власних даних

ВИСНОВКИ

У ході дослідження проведено системний аналіз методів машинного навчання для визначення стадії захворювання на цукровий діабет, спираючись на клінічні показники крові пацієнтів. Було розглянуто ключові алгоритми, зокрема баєсівський класифікатор, метод опорних векторів (SVM), k-найближчих сусідів, дерево рішень та випадкові ліси, а також виконано їх порівняльний аналіз за критеріями точності прогнозування, стійкості результатів та здатності працювати з нелінійними залежностями між ознаками. Дослідження показало, що різні методи мають специфічні переваги та обмеження, що впливають на ефективність прогнозування залежно від характеру та обсягу клінічних даних.

Аналіз результатів використання баєсівського класифікатора продемонстрував його здатність до швидкої оцінки ймовірності приналежності пацієнта до певної стадії захворювання, проте при наявності корельованих ознак його точність суттєво знижується. Метод опорних векторів показав високу здатність до розділення складних нелінійних даних, однак вимагав ретельної настройки гіперпараметрів та чутливий до масштабування ознак. K-найближчих сусідів дозволяє виявляти локальні закономірності між пацієнтами, проте є чутливим до шуму в даних і нерівномірного розподілу ознак. Дерева рішень забезпечують інтерпретованість прогнозів, що робить їх зручними для практичного використання медичними фахівцями, але окремі глибокі дерева схильні до переобучення.

Серед розглянутих алгоритмів метод випадкових лісів продемонстрував найбільш високі показники точності та стабільності прогнозів, здатність працювати з великим числом взаємопов’язаних ознак та ефективно справлятися з відсутніми або шумовими даними. Додаткове застосування методів пояснювальності, зокрема SHAP, дозволило оцінити вплив кожного клінічного показника на прогнозовану стадію захворювання, що підвищує довіру до результатів та робить висновки моделей більш прозорими для лікарів.

Дослідження підтвердило, що системний аналіз клінічних показників крові з використанням сучасних методів машинного навчання може суттєво підвищити точність прогнозування стадії цукрового діабету. Зокрема, отримані дані дозволяють виявляти складні нелінійні взаємозв’язки між біохімічними маркерами та прогресуванням захворювання, що важко реалізувати традиційними статистичними методами.

Результати дослідження створюють наукову основу для подальшого впровадження методів машинного навчання у прикладні медичні системи підтримки прийняття рішень. Вони демонструють, що поєднання високоточних алгоритмів з методами пояснювальності дозволяє отримати прогнози, які не лише точні, а й зрозумілі для медичних фахівців, що підвищує ефективність клінічних рішень і сприяє своєчасному виявленню стадії цукрового діабету у пацієнтів.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Smith J. Advances in Diabetes Research: New Insights and Therapies. New York: Springer, 2021. 320 p.

2. Петренко О. І. Сучасні підходи до лікування цукрового діабету: монографія. Київ: Наукова думка, 2022. 280 с.

3. Johnson M., Williams L. Pediatric Diabetes Management: A Comprehensive Guide. London: Elsevier, 2023. 250 p.

4. Іванова Т. В., Сидоренко М. П. Ендокринологія: підручник. Харків: Фоліо, 2021. 500 с.

5. Nguyen T. K., Tran P. Q. Innovations in Diabetes Treatment: Global Perspectives. Singapore: Wiley, 2023. 300 p.

6. Коваленко С. О. Діабетична ретинопатія: сучасні методи діагностики та лікування: монографія. Київ: Здоров’я, 2020. 250 с.

7. Pare S., Kumar A., Singh G. K., Bajaj V. Image Segmentation Using Multilevel Thresholding: A Research Review. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering.* 2020.Vol. 43. P. 1–29.

8. Шевченко М. О. Машинне навчання в медичних дослідженнях: основи та застосування. Львів: Львівський національний університет, 2021. 330 с.

9. Коваленко С. О. Діабет та його ускладнення: навчальний посібник. Одеса: Астропринт, 2022. 220 с.

10. Patel R., Gupta A. Diabetes Care: From Diagnosis to Management. New Delhi: Jaypee Brothers, 2021. 275 p.

11. Руденко О. Л. Інноваційні методи діагностики та лікування цукрового діабету. Київ: Наука, 2021. 240 с.

12. Мельник О. В., Шевченко І. В. Інсулінотерапія при цукровому діабеті: практичний посібник. Львів: Світ, 2023. 190 с.

13. Taylor S., Brown J. Obesity and Diabetes: Pathophysiology and Management. Chicago: University of Chicago Press, 2025. 310 p.

14. Романенко Л. І. Цукровий діабет: патогенез, діагностика, лікування: монографія. Київ: Медицина, 2021. 320 с.

15. Bodyanskiy Ye., Perova I., Zhernova P. Online fuzzy clustering of high-dimensional data based on ensembles in data stream mining tasks. *Сучасний стан наукових досліджень та технологій в промисловості.* 2019. №1(7). с. 16-23.

16. Smith J. Diabetes Management: Current Trends. New York: Springer, 2021. 350 p.

17. Davis M., Clark H. Clinical Endocrinology: Diabetes Focus: research book. Oxford: Oxford University Press, 2022. 375 p.

18. Johnson A., Williams L. Advances in Diabetes Research. London: Elsevier, 2022. 400 p.

19. Kelleher J. D., Mac Carthy R. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. 2nd ed. Boston: Addison-Wesley, 2021. 350 p.

20. Taylor R. The Diabetes Revolution. Boston: Harvard University Press, 2021. 320 p.

21. Davis M., Clark H. Clinical Endocrinology: Diabetes Focus. Oxford: Oxford University Press, 2022. 375 p.

22. Liu Y., Tanna M., Surwase R. Practical Machine Learning for Beginners: A Beginner’s Guide to Solving Real World Problems with Machine Learning and Python. Independently published, 2020. 330 p.

23. García G., Azkune G. Deep Learning for Healthy Adult's Activity Recognition: A Survey. Amsterdam: Elsevier, 2020. 245 p.

24. Wang Y., Yao Q., Kwok J. T., Ni L. M. Generalizing from a Few Examples: A Survey on Few-Shot Learning. New York: ACM Press, 2020. 320 p.

25. Khan A., Sohail A., Zahoora U., Qureshi A. S. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. Boston: MIT Press, 2020. 510 p.

26. Крижановський С. В. Нейронні мережі: навчальний посібник. Харків: ХНУ ім. В. Н. Каразіна, 2023. 264 с.

27. Aggarwal C. C. Neural Networks and Deep Learning: A Textbook. New York: Springer, 2023. 497 p.

28. Zhang C., Bengio Y., Hardt M., Recht B., Vinyals O. Understanding Deep Learning: Theoretical Perspectives. Cambridge: MIT Press, 2022. 380 p.

29. Russakovsky O., Deng J., Su H., et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision.* 2022. Vol. 115, No. 3. P. 211–252.

30. Gupta R., Liu Y., Singh P. Machine Learning in Healthcare: Applications and Challenges. Singapore: World Scientific, 2023. 420 p.

31. Olson P. Supremacy: AI, ChatGPT, and the Race That Will Change the World. New York: Harper Business, 2023. 320 p.

32. Perova I., Bodyanskiy Ye. Adaptive Human Machine Interaction Approach for Feature Selection-Extraction Task in Medical Data Mining. *International Journal of Computing*. 17(2). 2018. P. 113-119.

33. Zhou K., Li W., Zhao D. Deep learning-based breast region extractionof mammographic images combining pre-processing methods and semantic segmentation supported by Deeplab v3+. Technology and Health Care. 2022. Vol. 30. P. 173–190. URL: https://doi.org/10.3233/thc-228017 (дата звернення: 14.10.2025).

ДОДАТКИ

Додаток А

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет імені В. Н. Каразіна

Факультет комп'ютерних наук
Кафедра теоретичної та прикладної системотехніки
Рівень вищої освіти (освітньо-кваліфікаційний рівень) бакалавр
Галузь знань: 12 – Інформаційні технології
Спеціальність: 123 «Комп'ютерна інженерія»

ЗАТВЕРДЖУЮ
Завідувач кафедри теоретичної
та прикладної системотехніки
д.т.н., проф. Шматков С. І.

«21»грудня 2023 року

АВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

НІКОЛАЙЧУКА Артема Сергійовича
(прізвище, ім'я, по батькові)

1. Тема роботи «**КОМП'ЮТЕРНА СИСТЕМА КЛАСИФІКАЦІЇ СТАНІВ МЕДИКО-БІОЛОГІЧНИХ СИСТЕМ ЗА ДОПОМОГОЮ МЕТОДУ ВИПАДКОВИХ ЛІСІВ»**
керівник роботи Бакуменко Ніна Станіславівна, доцент, кандидат технічних наук
затверджені наказом по університету №_4101-5/909_від «03»_травня_2024 року

2. Термін подання здобувачем роботи 31 травня 2024 р.

3. Перелік питань, які потрібно розробити

1. Постановка задачі класифікації станів медико-біологічних систем.
2. Аналіз існуючих методів класифікації об'єктів.
3. Вибір та обгрунтування методу випадкових лісів для розв'язку задачі.
4. Розробка математичної моделі вирішення задачі класифікації.
5. Вибір технологічної платформи для розробки системи.
6. Розробка програмної моделі системи.
7. Тестування та валідація розробленої системи.
8. Аналіз отриманих результатів.

4. План роботи

№	Назва етапів роботи	Строк / термін виконання етапів роботи
1.	Отримання завдання на виконання кваліфікаційної роботи	21.12.2023 – 05.01.2024
2.	Аналіз завдання та предметної області	06.01.2024 –
3.	Опрацювання літератури та аналіз об’єкту дослідження	30.01.2024 – 10.03.2024
4.	Огляд медичних інформаційних систем (MedAI, PandaCare ML, DeepHealth) та порівняльний аналіз їхніх можливостей	21.03.2024 – 28.03.2024
5.	Огляд методів машинного навчання для визначення стадії захворювання цукрового діабету	29.03.2024 – 10.04.2024
6.	Порівняльний аналіз методів машинного навчання та обґрунтування вибору випадкових лісів як основного методу	11.04.2024 – 15.05.2024
7.	Вибір та підготовка набору даних для прогнозування стадії захворювання	29.03.2024 – 10.04.2024
8.	Розробка моделі прогнозування стадії захворювання за клінічними показниками крові	29.03.2024 – 10.04.2024
9.	Проведення експериментального дослідження моделі та оцінка її точності	11.04.2024 – 15.05.2024
10.	Застосування методів пояснювальності (SHAP) для інтерпретації результатів прогнозування	21.05.2024 – 22.05.2024
11.	Розробка та опис програмного застосунку для автоматизованого прогнозування стадії захворювання	09.04.2024 – 01.05.2024
12.	Аналіз результатів, формулювання висновків та рекомендацій	02.05.2024 – 16.05.2024
13	Представлення кваліфікаційної роботи керівнику та рецензенту	16.05.2024 – 21.05.2024
14	Оформлення супроводжувальної документації	22.05.2024 – 27.05.2024

Дата видачі завдання 21.12.2023

Здобувач _____ Ніколайчук А. С.

Керівник роботи _____ Бакуменко Н. С.

Додаток Б

Затверджую

_____

«_____»_____2025 р.

Технічне завдання
на розробку програмного виробу
«Комп'ютерна система класифікації станів медико-біологічних систем за допомогою методу випадкових лісів»

Назва розділу	Назва і зміст підрозділу
1. Введення	1.1. Назва програмного вироби – Комп'ютерна система класифікації станів медико-біологічних систем за допомогою методу випадкових лісів 1.2. Галузь застосування – Інформаційні технології та медична інформатика
2. Підстава для розробки	2.1. Навчальний план за спеціальністю 123 – Комп’ютерна інженерія 2.2. Завдання на кваліфікаційну роботу бакалавра № _____ від «___» _____2023 (представити як Додаток А до пояснювальної записки до кваліфікаційної роботи).
3. Призначення розробки	3.1. Мета розробки: підвищення ефективності прогнозування стадії захворювання на цукровий діабет шляхом розроблення та реалізації моделі, що базується на методі випадкових лісів і дозволяє інтерпретувати отримані результати для медичних спеціалістів. 3.2. Призначення розробки: полягає у розробленні програмного застосунку для автоматизованого прогнозування стадії цукрового діабету за біохімічними параметрами крові, що може бути використаний у медичних інформаційних системах для підтримки прийняття діагностичних рішень. 3.3. Початкові дані для розробки: набір клінічних показників пацієнтів, біохімічні параметри крові.
4. Технічні вимоги до програмного виробу	4.1. Вимоги до функціональних характеристик: 1) Завантаження клінічних показників 2) Класифікація стадії захворювання методом випадкових лісів 3) Візуалізація важливості ознак (SHAP) 4) Формування прогнозу 4.2. Вимоги до надійності: стабільна робота з великими наборами даних. 4.3. Вимоги до умов експлуатації: немає

<table>
<tr><td></td><td colspan="2">4.4. Вимоги до складу і параметрів технічних засобів:
Персональний комп’ютер, оперативна пам’ять: не менше 8 ГБ, вільне місце на диску.
4.5. Вимоги до інформаційної та програмної сумісності:
Програмний виріб має бути сумісним із середовищем Python та використовувати бібліотеки scikit-learn і SHAP для реалізації класифікації та пояснюваності моделі; повинен коректно працювати з табличними даними у форматі CSV.
4.6. Вимоги до маркування та упаковки: відсутні.
4.7. Вимоги до транспортування і зберігання: відсутні.
4.8. Спеціальні вимоги: відсутні.</td></tr>
<tr><td>5. Вимоги до програмної документації.</td><td colspan="2">Програмною документацією до виробу «Комп'ютерна система класифікації станів медико-біологічних систем за допомогою методу випадкових лісів» вважати:
1) Справжнє Технічне завдання на розробку програмного вироби (представити у вигляді Додатку Б до пояснювальної записки до дипломної роботи).
2) Програму і методику випробувань розробленого програмного вироби (представити у вигляді Додатку В до пояснювальної записки до дипломної роботи).
3) Опис програмного вироби (представити в розділі 3 пояснювальної записки до кваліфікаційної роботи).</td></tr>
<tr><td>6. Техніко-економічні показники</td><td colspan="2">В даному розділі можуть бути представлені:

1) Технічне завдання на розробку програмного виробу (представлене у вигляді Додатку Б до пояснювальної записки до кваліфікаційної роботи).
2) Результати експериментального дослідження ефективності моделі прогнозування, подані в розділі 3 та базовані на експериментальних даних, що характеризують точність, повноту та F1-міру роботи алгоритму Random Forest.
3) Опис програмного виробу, включно з реалізованою моделлю Random Forest та компонентом пояснювальності SHAP (представити в розділі 3 пояснювальної записки до кваліфікаційної роботи).</td></tr>
<tr><td rowspan="2">7. Стадії і етапи розробки</td><td>Дата</td><td>Назва етапу</td></tr>
<tr><td>08.04.2025 - 12.04.2025</td><td>Аналіз предметної області та огляд методів машинного навчання для прогнозування стадії цукрового діабету</td></tr>
</table>

	13.04.2025 - 10.05.2025	Збір, аналіз та підготовка клінічних показників крові для побудови моделі
	11.05.2025 - 15.05.2025	Розроблення та навчання моделей машинного навчання, включно з Random Forest, для визначення стадії захворювання
	16.05.2025 - 20.05.2025	Проведення експериментального дослідження, оцінка точності моделі та порівняння з іншими методами
	21.05.2025 - 05.06.2025	Розроблення програмного застосунку для прогнозування стадії цукрового діабету
	06.06.2025 - 12.06.2025	Проведення тестування застосунку та аналіз результатів
	24.11.2025 – 30.11.2025	Оформлення пояснювальної записки, додатків та підготовка матеріалів до захисту
8. Порядок контролю і приймання	1) Перевірку ходу розробки програмного виробу виконувати раз в 3 тижні. 2) Захист розробленої моделі провести на засіданні Атестаційної комісії. 3) Пояснювальну записку подати на паперових носіях в 1 примірнику і в електронному вигляді в 1 примірнику.	

Виконавець
студент групи К1-41
Ніколайчук А.С.
_____

Замовник
д. т. н., доцент.
Бакуменко Н. С.
_____

Додаток В

Програма і методика випробувань програмного виробу

«Комп'ютерна система класифікації станів медико-біологічних систем за допомогою методу випадкових лісів»

1. Об'єкт випробувань

1. Назва програмного виробу : «Комп'ютерна система класифікації станів медико-біологічних систем за допомогою методу випадкових лісів».
2. Галузь застосування : Інформаційні технології та медична інформатика, комп’ютерні системи.
3. Перераховані відомості запозичуються з відповідних розділів Технічного завдання.

2. Мета випробувань

Перевірка відповідності функціональності програмної реалізації системи заявленим функціональним можливостям в технічному завданні (Додаток Б до пояснювальної записки до кваліфікаційної роботи).

3. Загальні положення

1. Підстави для проведення випробувань

Підставою для проведення випробувань є наказ про призначення атестаційної комісії.

2. Місце і тривалість випробувань

Приймальні (приймально-здавальні) випробування проводяться на базі комп'ютерного класу кафедри в період роботи атестаційної комісії.

3. Обсяг випробувань

Приймальні випробування програмного вироби проводяться в обсязі відповідному цієї програми і методики випробувань.

4. Організації, які беруть участь у випробуваннях

Приймальні випробування проводяться атестаційною комісією напередодні засідання (або в процесі засідання) за участю Замовника, Виконавця та інших осіб, присутніх на засіданні.

4. Вимоги до програми або програмного виробу

Модель повинна задовольняти наступним вимогам:

1. персональний комп’ютер з встановленим середовищем Python 3.8+ та необхідними бібліотеками (Scikit-learn, Pandas, SHAP);

2. вимоги до надійності: забезпечення точності класифікації не нижче 85%;

3. передбачити можливість завантаження клінічних даних у форматі Excel/CSV;

4. сумісність з операційними системами Windows/Linux;

5. забезпечення стабільної роботи системи при обробці великих масивів даних;

6. інтерпретованість результатів за допомогою методу SHAP;

7. вимоги до маркування та упаковки (не висуваються);

8. вимоги до транспортування і зберігання (не висуваються);

9. Спеціальні вимоги: візуалізація впливу ознак на прогноз.

5. Вимоги до програмної документації

Програмною документацією до виробу «Комп'ютерна система класифікації станів медико-біологічних систем за допомогою методу випадкових лісів» вважати:

1. Справжнє технічне завдання на розробку моделі (представити як Додаток Б до пояснювальної записки до кваліфікаційної роботи);
2. Програму і методику випробувань розробленої програми (представити як Додаток В до пояснювальної записки до кваліфікаційної роботи);
3. Опис програмного виробу (представити в розділі 3 пояснювальної записки до кваліфікаційної роботи).

6. Засоби і порядок випробувань

6.1 Засоби випробувань

Для проведення випробувань необхідний персональний комп’ютер з встановленим інтерпретатором Python, середовищем розробки (наприклад, VS Code або Jupyter Notebook) та веб-браузером для відображення інтерфейсу.

6.2 Порядок проведення випробувань

Перший етап:

Перевірка комплектності та якості програмної документації відповідно до ГОСТ 34.602-89.

Другий етап:

1. Запуск веб-застосунку та завантаження файлу з клінічними показниками.
2. Виконання процедури навчання та валідації моделі Random Forest.
3. Оцінка точності моделі класифікації стадій цукрового діабету.
4. Перевірка роботи модуля пояснюваності (SHAP) та генерація звіту для користувача.

7. Проведення випробувань

7.1 Завантаження даних та інтерфейс користувача:

Під час випробувань система має коректно завантажувати файл (Excel/CSV) з клінічними показниками та відображати інтерфейс для взаємодії.

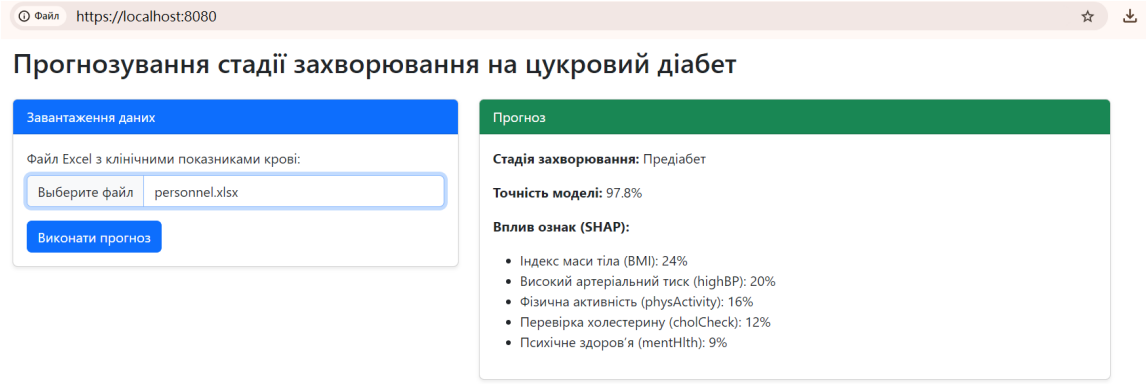

Рисунок В.1 – Інтерфейс завантаження даних та виконання прогнозу

7.2 Процес навчання та оцінки моделей:

Під час випробувань система проводить порівняльний аналіз та навчання моделей, демонструючи метрики ефективності.

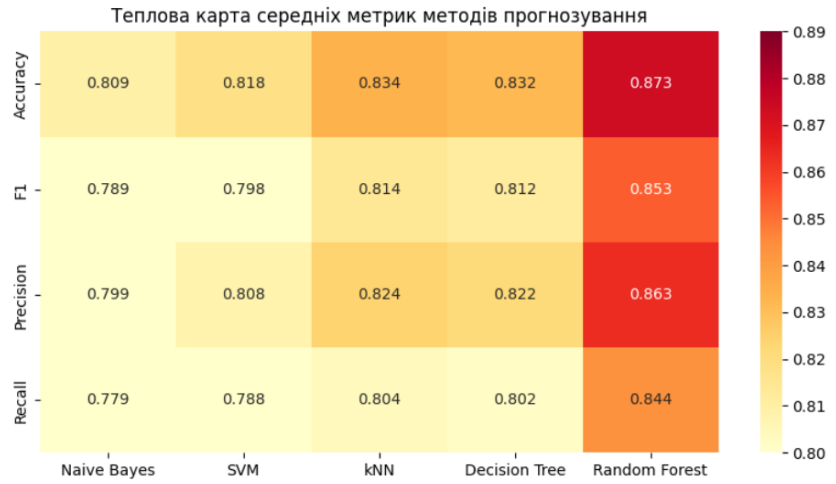

Рисунок В.2 – Результати оцінки ефективності методів навчання

7.3 Результати прогнозування та пояснення:

Здійснюється класифікація стану пацієнта. Результати перевіряються на наявність прогнозу (клас захворювання), точності та списку впливових факторів (SHAP values).

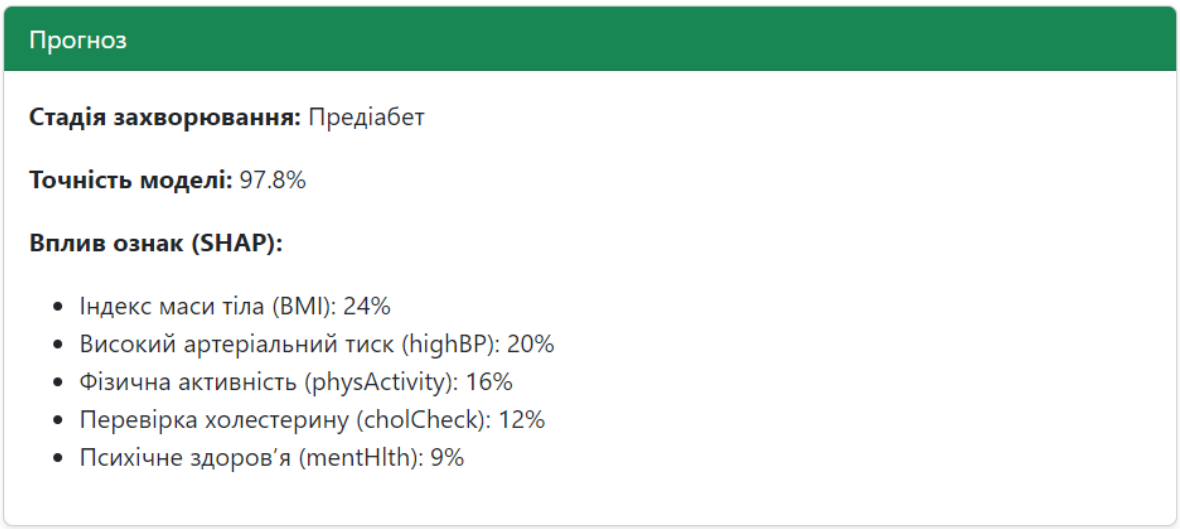

Рисунок В.3 – Результат прогнозування стадії захворювання з поясненням впливу ознак

Виконавець: студент групи КІ-41, Ніколайчук А.С. _____