

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Навчально-науковий інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

До захисту допущено:
Завідувач кафедри
_____ Оксана ТИМОЩУК
« ____ » _____ 2025 р.

Дипломна робота
на здобуття ступеня бакалавра
за освітньо-професійною програмою «Системний аналіз і управління»
спеціальності 124 «Системний аналіз»
на тему: «Методи машинного навчання для прогнозування ризику
виникнення захворювання Альцгеймера»

Виконала:
студентка IV курсу, групи КА-14
Петренко Карина Ігорівна

Керівник: _____
Завідувач кафедри ММСА, к.т.н., доцент
Тимошук Оксана Леонідівна

Консультант з економічного розділу: _____
доцент, к.е.н. Рощина Надія Василівна _____

Консультант з нормоконтролю: _____
доцент, к. ф.-м. н. Статкевич Віталій Михайлович _____

Рецензент: _____
доцент кафедри системного проектування,
к.т.н., с.н.с. Кисельов Геннадій Дмитрович _____

Засвідчую, що у цій дипломній роботі
немає запозичень з праць інших авторів
без відповідних посилань.
Студентка _____

Київ – 2025 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Навчально-науковий інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 124 «Системний аналіз»

Освітньо-професійна програма «Системний аналіз і управління»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Оксана ТИМОЩУК

«__» _____ 2025 р.

ЗАВДАННЯ

на дипломну роботу студентці

Петренко Карині Ігорівні

1. Тема роботи «Методи машинного навчання для прогнозування ризику виникнення захворювання Альцгеймера», керівник роботи Тимощук Оксана Леонідівна, к.т.н., доцент, затверджені наказом по університету від «26» травня 2025 р. № 1759-с

2. Термін подання студентом роботи: _____

3. Вихідні дані до роботи: набір медичних даних та комплексної інформації про пацієнтів.

4. Зміст роботи: дослідження сучасних методів машинного навчання для прогнозування ризику виникнення захворювання Альцгеймера, проаналізувати ефективність моделей класифікації, оцінити якість прогнозування за допомогою обраних метрик та сформулювати висновки щодо доцільності застосування обраних підходів.

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо): презентація.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Економічний	доцент, к.е.н. Рощина Н.В.		

7. Дата видачі завдання 14.04.2025

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Затвердження теми ДР	03.02.2025-19.02.2025	Виконано
2	Ознайомлення зі структурою БДР згідно з Положенням про державну атестацію студентів НТУУ «КПІ ім. І. Сікорського»	20.02.2025-03.03.2025	Виконано
3	Ознайомлення з ДСТУ 3008:2015	04.03.2025-10.03.2025	Виконано
4	Проведення дослідження за темою БДР під керівництвом керівника	11.03.2025-04.04.2025	Виконано
5	Завершення роботи над першим варіантом частини БДР	05.04.2025-13.04.2025	Виконано
6	Проведення роботи над експериментальною частиною БДР та програмною реалізацією	14.04.2025-18.05.2025	Виконано
7	Виконання функціонально-вартісного аналізу розробленого програмного забезпечення	19.05.2025-21.05.2025	Виконано
8	Підготовка презентації БДР	22.05.2025-01.06.2025	Виконано
9	Оформлення БДР та аналіз отриманих результатів	02.06.2025-15.06.2025	Виконано

Студентка

Карина ПЕТРЕНКО

Керівник

Оксана ТИМОЩУК

РЕФЕРАТ

Дипломна робота складається зі 143 сторінок, містить 27 рисунків, 12 таблиць, 2 додатки та 33 джерела літератури.

ЗАХВОРЮВАННЯ АЛЬЦГЕЙМЕРА, МАШИННЕ НАВЧАННЯ, КЛАСИФІКАЦІЯ, ПРОГНОЗУВАННЯ, НЕЙРОДЕГЕНЕРАТИВНІ ЗАХВОРЮВАННЯ.

Об'єктом дослідження стала хвороба Альцгеймера як одне з найбільш поширених нейродегенеративних захворювань.

Метою дослідження є розробка та оцінка ефективності моделей машинного навчання для прогнозування ризику виникнення захворювання Альцгеймера на основі медичних та когнітивних ознак.

Проблема має високу актуальність, оскільки рання діагностика хвороби Альцгеймера має вирішальне значення для своєчасного втручання та покращення якості життя пацієнтів. Застосування алгоритмів машинного навчання є перспективним підходом у створенні інтелектуальних систем підтримки прийняття рішень у сфері медицини.

У ході роботи було реалізовано низку моделей машинного навчання, зокрема Decision Tree, Random Forest, XGBoost, Support Vector Machine та Stacking Ensemble. Кодування та експерименти здійснювались мовою програмування Python у середовищі Jupyter Notebook. Для оцінювання якості моделей використовувались основні метрики класифікації: Accuracy, Precision, Recall, F1-score та AUC (Area Under Curve). Здійснено аналіз п'яти моделей машинного навчання, порівняно їх характеристики і відібрано найбільш доцільні для задачі прогнозування.

У рамках майбутніх досліджень планується впровадження додаткових методів підвищення точності моделей, таких як використання більшої кількості медичних показників, інтеграції часових змінних для прогнозування динаміки захворювання, а також балансування класів.

ABSTRACT

The Bachelor thesis contains 143 pages, 27 figures, 12 tables, 2 appendices, 33 references.

ALZHEIMER'S DISEASE, MACHINE LEARNING, CLASSIFICATION, FORECASTING, NEURODEGENERATIVE DISEASES.

The object of the study is Alzheimer's disease as one of the most common neurodegenerative diseases.

The aim of the study is to develop and evaluate the effectiveness of machine learning models for predicting the risk of Alzheimer's disease based on medical and cognitive characteristics.

The problem is highly relevant, since early diagnosis of Alzheimer's disease is crucial for timely intervention and improving the quality of life of patients. The use of machine learning algorithms is a promising approach in creating intelligent decision support systems in the field of medicine.

During the study, a number of machine learning models were implemented, including Decision Tree, Random Forest, XGBoost, Support Vector Machine and Stacking Ensemble. Coding and experiments were carried out in the Python programming language in the Jupyter Notebook environment. The main classification metrics were used to assess the quality of the models: Accuracy, Precision, Recall, F1-score and AUC (Area Under Curve). During the study, five models were implemented to predict the risk of Alzheimer's disease. The accuracy and efficiency of the models were analyzed based on the obtained metrics. The strengths and limitations of each model were also identified and recommendations for their further improvement were provided.

In the framework of future studies, it is planned to implement additional methods to increase the accuracy of the models, such as using a larger number of medical indicators, integrating time variables to predict the dynamics of the disease, and balancing classes.

ЗМІСТ

ВСТУП	9
РОЗДІЛ 1 ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ.....	11
1.1 Постановка задачі.....	11
1.2 Актуальність досліджуваної проблеми.....	12
1.3 Етіологія та фактори ризику	17
1.4 Діагностика.....	24
1.4.1 Традиційні методи діагностики	24
1.4.2 Передові методики діагностики.....	28
1.4.3 Роль машинного навчання у діагностиці	32
1.5 Висновки до розділу 1.....	33
РОЗДІЛ 2 МАТЕМАТИЧНІ МЕТОДИ ТА МОДЕЛІ ЯК ІНСТРУМЕНТ ДІАГНОСТИКИ ЗАХВОРЮВАННЯ АЛЬЦГЕЙМЕРА	35
2.1 Основні поняття і термінологія машинного навчання	35
2.2 Математичні основи машинного навчання.....	40
2.3 Алгоритми машинного навчання	44
2.3.1 Decision Tree	44
2.3.2 Random Forest	50
2.3.3 XGBoost.....	55
2.3.4 Support Vector Machine	61
2.3.5 Stacking Ensemble	66
2.4 Висновки до розділу 2.....	71

РОЗДІЛ 3 РЕАЛІЗАЦІЯ ПРОГРАМНОГО ПРОДУКТУ ТА АНАЛІЗ РЕЗУЛЬТАТІВ	73
3.1 Огляд набору даних	73
3.2 Попередня обробка даних.....	77
3.3 Метрики оцінки отриманих результатів.	83
3.4 Навчання моделей і аналіз отриманих результатів	86
3.4.1 Decision Tree	87
3.4.2 Random Forest	89
3.4.3 XGBoost.....	91
3.4.4 Support Vector Machine.....	93
3.4.5 Stacking Ensemble	95
3.5 Висновки до розділу 3.....	97
РОЗДІЛ 4 ФУНКЦІОНАЛЬНО-ВАРТІСНИЙ АНАЛІЗ ПРОГРАМНОГО ПРОДУКТУ	99
4.1 Постановка задачі проектування.....	100
4.2 Обґрунтування функцій програмного продукту	100
4.3 Обґрунтування системи параметрів програмного продукту	104
4.4 Аналіз експертного оцінювання параметрів.....	107
4.5 Аналіз рівня якості варіантів реалізації функцій	110
4.6 Економічний аналіз варіантів розробки ПП.....	111
4.7 Вибір кращого варіанту ПП техніко-економічного рівня.....	117
4.8 Висновки до розділу 4.....	118
ВИСНОВКИ	119
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ	121
ДОДАТКИ	125

Додаток А. Програмна реалізація алгоритмів	125
Додаток Б. Презентація.....	134

ВСТУП

Захворювання Альцгеймера є однією з найпоширеніших форм деменції, що вражає мільйони людей у світі, особливо людей похилого віку. Це прогресуюче нейродегенеративне захворювання не лише призводить до значного зниження когнітивних функцій, а й має серйозні соціальні та економічні наслідки для пацієнтів, їхніх родин і систем охорони здоров'я загалом. У зв'язку з цим, рання діагностика ризику виникнення хвороби Альцгеймера має вирішальне значення для своєчасного втручання та поліпшення якості життя пацієнтів.

Сучасна медицина стикається з необхідністю обробки великих обсягів медичних даних, що відкриває широкі можливості для використання інструментів штучного інтелекту, зокрема методів машинного навчання (ML). Ці методи здатні виявляти приховані закономірності в медичних даних і будувати прогностичні моделі з високою точністю, що робить їх надзвичайно перспективними для діагностики та прогнозування нейродегенеративних захворювань, зокрема хвороби Альцгеймера.

Актуальність використання машинного навчання в діагностиці хвороби Альцгеймера полягає в його здатності виявляти ранні ознаки нейродегенеративних змін, що дозволяє завчасно реагувати, уповільнювати прогрес хвороби та покращувати якість життя пацієнтів. Завдяки потужним аналітичним можливостям, алгоритми машинного навчання здатні обробляти великі обсяги різномірних медичних даних, виявляти приховані закономірності та формувати прогностичні моделі з високою точністю.

Разом із тим, ефективне застосування алгоритмів машинного навчання для діагностики хвороби Альцгеймера є непростим завданням, яке потребує комплексного підходу: від ретельного аналізу вхідних даних до вибору оптимальних моделей та правильного тлумачення отриманих результатів. Це

обумовлено багатфакторною природою захворювання та неоднорідністю медичних ознак, які можуть впливати на діагностику.

У межах дослідження було поставлено за мету проаналізувати ефективність декількох алгоритмів машинного навчання для прогнозування ризику розвитку хвороби Альцгеймера, зокрема Decision Tree, Random Forest, XGBoost, Support Vector Machine та Stacking Ensemble. Моделі було реалізовано за допомогою мови програмування Python у середовищі Jupyter Notebook. Було проведено аналіз точності та ефективності побудованих моделей на основі отриманих метрик. Також було виявлено сильні сторони й обмеження кожної з моделей і надано рекомендації щодо їх подальшого покращення. Основою для розробки став набір медичних даних Alzheimer's Disease Dataset, що містить комплексну інформацію про пацієнтів, зокрема: демографічні дані, фактори способу життя, медичну історію, клінічні вимірювання, оцінки когнітивного й функціонального стану, а також наявність симптомів.

Перший розділ дипломної роботи присвячено аналізу предметної області, зокрема етіології хвороби Альцгеймера, сучасним методам діагностики та використанню ML у нейродегенеративній медицині. У другому розділі детально розглянуто основні поняття машинного навчання, описано застосовані алгоритми, їхні математичні основи, переваги та недоліки. Третій розділ охоплює процес попередньої обробки даних, побудову моделей та аналіз їх результатів. Четвертий розділ присвячений функціонально-вартісному аналізу розробленого програмного продукту.

Таким чином, результати цього дослідження підтверджують доцільність і ефективність використання алгоритмів машинного навчання для прогнозування хвороби Альцгеймера. Запропоновані підходи мають потенціал для інтеграції в системи підтримки прийняття медичних рішень, сприяючи ранньому виявленню захворювання, підвищенню точності діагностики та покращенню якості життя пацієнтів.

РОЗДІЛ 1 ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Постановка задачі

У межах даної дипломної роботи розглядається задача прогнозування ризику виникнення захворювання Альцгеймера за допомогою методів машинного навчання. Вхідними даними є медичний датасет, що містить інформацію про демографічні характеристики пацієнтів, фактори способу життя, анамнез, клінічні вимірювання, оцінки когнітивного та функціонального стану, а також симптоматику. Такі дані є структурованими та представлені у вигляді табличних записів, кожен з яких відповідає окремому пацієнту.

Метою є побудова та оцінка прогностичних моделей, які дозволяють класифікувати пацієнтів за наявністю або відсутністю ризику розвитку хвороби Альцгеймера. Задача формалізується як задача класифікації зі скінченною кількістю класів (наприклад: «ризик відсутній», «ризик наявний»). Процес навчання здійснюється на наборі даних, де для кожного прикладу відома цільова змінна — наявність або відсутність діагнозу.

Під час дослідження буде реалізовано п'ять моделей для прогнозування ризику виникнення захворювання Альцгеймера. На основі аналізу результатів буде проведено порівняння моделей за якістю класифікації, виявлено їх переваги та недоліки. Також розглядатимуться підходи до покращення моделей, зокрема шляхом налаштування гіперпараметрів, використання технік балансування класів та зниження розмірності даних.

У результаті мають бути побудовані та апробовані високоточні моделі машинного навчання, здатні ефективно прогнозувати ризик розвитку хвороби Альцгеймера, що може мати значну цінність для клінічної практики та систем ранньої діагностики.

1.2 Актуальність досліджуваної проблеми

Згідно з даними Всесвітньої організації охорони здоров'я (ВООЗ), хвороба Альцгеймера є найпоширенішою формою деменції, на яку припадає 60–70% усіх випадків. Наразі деменція посідає сьоме місце серед провідних причин смертності у світі та є однією з головних причин інвалідності й залежності літніх людей від допомоги інших. У 2019 році глобальні економічні витрати, пов'язані з деменцією, становили 1,3 трильйона доларів США, причому близько 50% цих витрат припадало на неофіційний догляд — зокрема з боку родичів і друзів, які в середньому приділяють цьому близько 5 годин щодня [1].

Крім того, щороку реєструється понад 5 мільйонів нових випадків хвороби Альцгеймера. Рівень захворюваності зростає з приблизно 1% серед людей віком 60–70 років до 6–8% серед осіб віком від 85 років і старше. Близько 10% людей старше 70 років мають виражені порушення пам'яті, і більш ніж у половини з них виявляються клінічні симптоми хвороби Альцгеймера. Тривалість перебігу захворювання зазвичай становить від 8 до 10 років, хоча в деяких випадках може варіюватися від 2 до 25 років після встановлення діагнозу [2]. Згодом прогнозується подальше зростання частоти цього захворювання у міру старіння населення планети, що підкреслює глобальний характер проблеми.

Наслідки хвороби Альцгеймера є багатовимірними: крім медичного та психологічного аспектів, вона спричиняє суттєві соціальні та економічні затрати. Пацієнти потребують паліативного догляду, що створює фінансове навантаження на родини, соціальні служби та систему охорони здоров'я загалом. Водночас рання діагностика та прогнозування ризику розвитку хвороби є важливою умовою для своєчасного медичного втручання, що може продовжити функціональну незалежність пацієнтів і підвищити якість їхнього життя. Зважаючи на складність клінічної діагностики на ранніх етапах,

важливого значення набуває впровадження сучасних цифрових технологій, зокрема методів машинного навчання, які здатні аналізувати великі обсяги медичних даних та виявляти приховані закономірності, що недоступні при традиційному аналізі. Це відкриває нові можливості для побудови ефективних інтелектуальних систем підтримки прийняття медичних рішень у сфері нейродіагностики.

Важливо зазначити, що хвороба Альцгеймера суттєво впливає як на життя окремих пацієнтів, так і на добробут їхніх родин. Люди, які страждають на цю недугу, зазнають поступового зниження когнітивних функцій, що призводить до втрати незалежності, зниження якості життя, соціальної ізоляції та глибоких психологічних змін. Пацієнти часто стикаються з проблемами самообслуговування, комунікації, просторової орієнтації, що унеможливорює їх самостійне існування на пізніх стадіях захворювання. Цей тягар лягає також на близьке оточення, яке змушене брати на себе функції догляду, супроводу та забезпечення безпеки хворого.

Крім соціального та психологічного впливу, хвороба Альцгеймера спричиняє вагоме економічне навантаження як на родини, так і на систему охорони здоров'я та національну економіку в цілому. Витрати, пов'язані з доглядом за хворими, охоплюють як прямі (медичне обслуговування, медикаменти, госпіталізація, довготривале перебування в спеціалізованих закладах), так і непрямі витрати (втрата працездатності пацієнтів).

За результатами глобального макроекономічного моделювання, що враховує втрати трудових і капітальних ресурсів унаслідок смертності, захворюваності та неформального догляду, загальний економічний тягар захворювання Альцгеймера та інших деменцій у період з 2020 по 2050 рік становитиме приблизно 14 513 мільярдів доларів США, що еквівалентно щорічному податку у розмірі 0,421% на світовий ВВП протягом 2020–2050 років. Це співставно з усіма глобальними витратами на охорону здоров'я у 2020 році [3]. Макроекономічний тягар захворювання Альцгеймера та інших

видів деменції з 2020 по 2050 рік виражений у мільярдах міжнародних доларів та у відсотках від загального ВВП зображено на рис. 1.1 та рис. 1.2 відповідно.

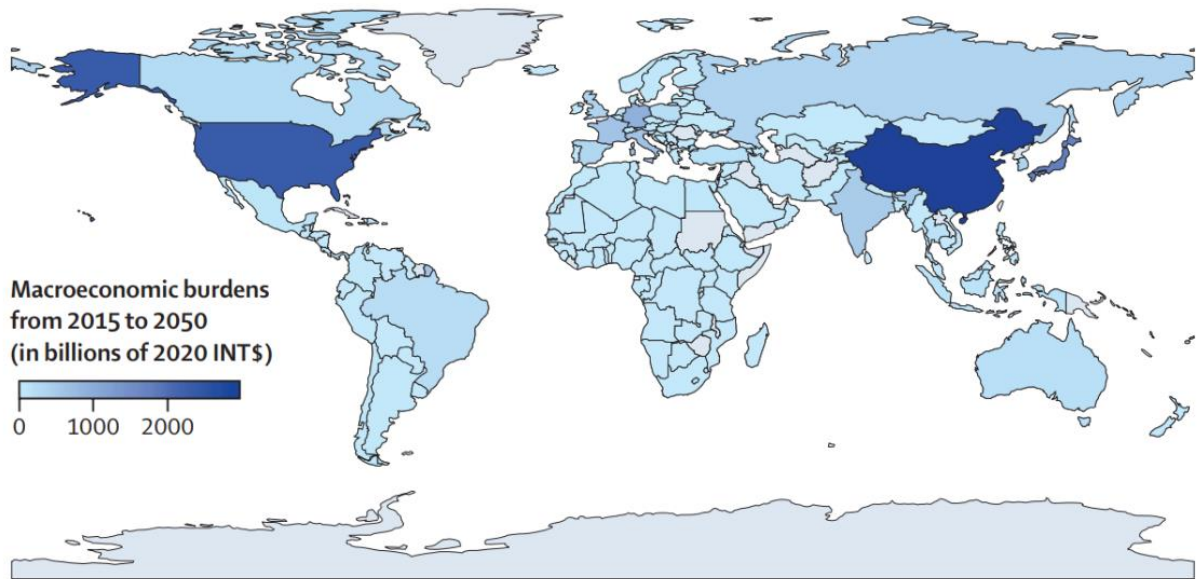


Рисунок 1.1 - Макроекономічний тягар захворювання Альцгеймера та інших видів деменції з 2020 по 2050 рік виражений у мільярдах міжнародних доларів [4]

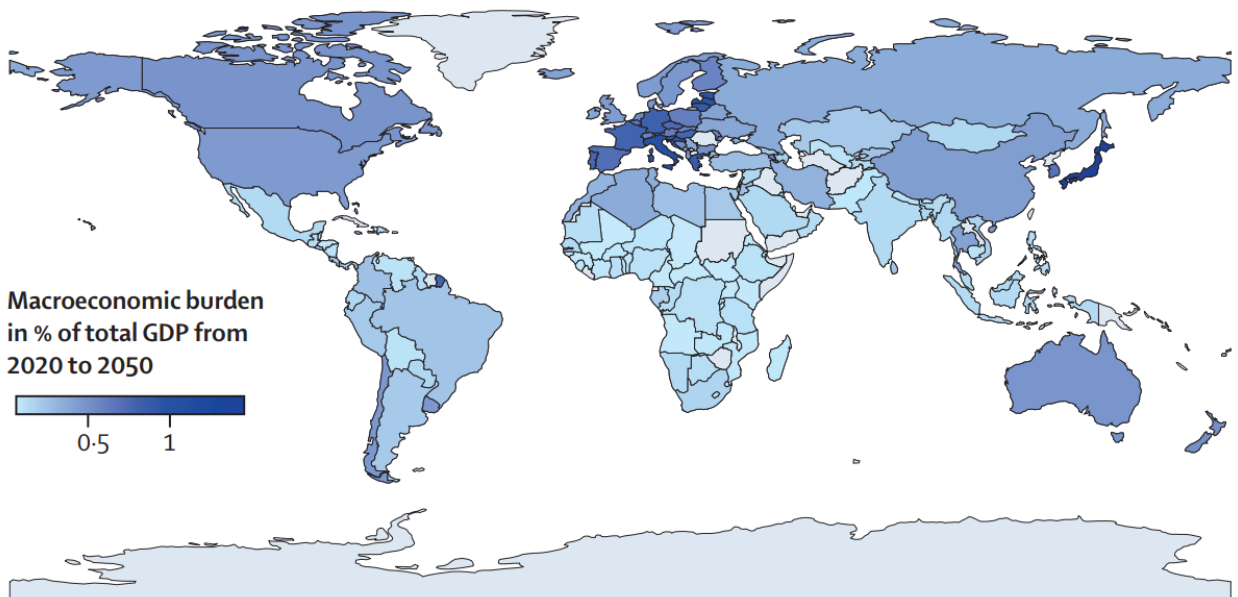


Рисунок 1.2 - Макроекономічний тягар захворювання Альцгеймера та інших видів деменції з 2020 по 2050 рік у відсотках від загального ВВП [5]

Раніше інші підходи до оцінки тягаря захворювання Альцгеймера та інших деменцій, зокрема методологія вартості захворювання, показали ще вищі оцінки: у 2015 році прямі витрати становили 0,65% світового ВВП, а витрати на неформальний догляд – 0,44%, у той час як у 2019 році ці показники досягли 0,76% і 0,75% відповідно. За прогнозами ВООЗ, до 2030 року загальна вартість захворювання Альцгеймера та інших видів деменцій може сягнути 2,8 трильйона доларів США з урахуванням зростання витрат на неформальний та формальний догляд, що вказує на стрімке зростання навантаження на систему охорони здоров'я, економіку та суспільство [6].

Така фінансова нестабільність обумовлена як зростанням чисельності хворих, так і тривалістю захворювання, що часто триває понад десятиліття. Отже, своєчасне прогнозування ризику та рання діагностика хвороби Альцгеймера є не лише клінічно доцільними, а й економічно обґрунтованими заходами, які можуть зменшити навантаження на державні бюджети та покращити розподіл ресурсів охорони здоров'я.

Ситуація в Україні загалом подібна до тієї, що спостерігається в Європі та світі. Хвороба Альцгеймера входить до переліку десяти основних захворювань, при яких пацієнти потребують паліативної допомоги. Водночас рівень смертності від цієї хвороби серед дорослого населення в Україні становить 0,3 випадка на 100 000 осіб. Для порівняння, у Великій Британії цей показник дорівнює 112, у США — 82, у Франції — 76, а в Німеччині — 51 на 100 000 населення. За кількістю смертей, спричинених хворобою Альцгеймера, Україна посідає 136 місце у світовому рейтингу [7].

Зважаючи на стрімке зростання поширеності хвороби Альцгеймера, яка є однією з найбільш поширених форм деменції, дослідження та розвиток новітніх методів діагностики та прогнозування цієї хвороби є надзвичайно важливими. Раннє виявлення захворювання дозволяє значно полегшити перебіг хвороби та запобігти важким когнітивним розладам, що знижують якість життя пацієнтів. Точна та своєчасна діагностика хвороби Альцгеймера може дати найкращий шанс підготуватися та спланувати майбутнє, а також

отримати лікування чи підтримку, які можуть допомогти. Комплексне розуміння етіології, факторів ризику, патофізіології та оптимальних стратегій лікування хвороби Альцгеймера є необхідним для розробки ефективних профілактичних і терапевтичних підходів, які можуть допомогти зберегти когнітивні функції на ранніх стадіях захворювання. Враховуючи складність хвороби Альцгеймера та її багатфакторну природу, вирішення цієї проблеми вимагає мультидисциплінарного підходу, що включає співпрацю медичних працівників, науковців та фахівців в галузі технологій. Методи машинного навчання відкривають нові можливості в аналізі медичних даних і прогнозуванні ризику розвитку хвороби Альцгеймера.

Інтеграція методів машинного навчання в медичну практику дозволяє не тільки покращити точність діагностики, а й розробити системи підтримки прийняття рішень для лікарів, що сприяють персоналізованому підходу до лікування та прогнозування хвороби Альцгеймера. Алгоритми машинного навчання здатні обробляти великі обсяги різноманітних медичних даних (демографічні характеристики, когнітивні оцінки, фактори способу життя та ін.), що значно підвищує точність прогностичних систем.

Отже, актуальність дослідження методів машинного навчання для прогнозування ризику виникнення хвороби Альцгеймера зростає у зв'язку з потребою в удосконаленні діагностичних та профілактичних заходів. Впровадження новітніх технологій у цю сферу дозволяє не тільки покращити якість життя пацієнтів, але й знизити економічне навантаження на систему охорони здоров'я та суспільство. Запровадження машинного навчання у нейродіагностику хвороби Альцгеймера надає нові можливості для більш ефективної боротьби з цим нейродегенеративним захворюванням та значно покращує прогнози щодо її лікування й управління.

1.3 Етіологія та фактори ризику

Етіологія хвороби Альцгеймера є складною та поліетіологічною. Розвиток цього нейродегенеративного захворювання пов'язаний із поєднаною дією генетичних, біохімічних, метаболічних, поведінкових та середовищних факторів. Серед основних етіологічних чинників виділяють вік, генетичну схильність, низький рівень освіти, травми голови, а також шкідливі звички та малорухливий спосіб життя. Розуміння етіологічних аспектів хвороби Альцгеймера має вирішальне значення для виявлення осіб із підвищеним ризиком, а також для розробки ефективних стратегій профілактики, раннього втручання та уповільнення прогресу захворювання. Саме тому дослідження взаємозв'язку між факторами ризику та появою клінічних симптомів є одним із ключових завдань сучасної нейронауки.

Незмінні фактори ризику захворювання Альцгеймера

Хвороба Альцгеймера має низку незмінних факторів ризику, які неможливо усунути або змінити шляхом профілактичних заходів. До таких факторів належать, насамперед, вік, генетична схильність та сімейний анамнез захворювання. Імовірність розвитку хвороби зростає з віком, ризик суттєво підвищується після 65 років і досягає піку у віці понад 85 років [8]. Генетичні чинники також відіграють істотну роль — зокрема, ген АР0Е істотно підвищує імовірність розвитку хвороби. Крім того, якщо в сімейному анамнезі вже мали місце випадки Альцгеймера, ризик для інших членів родини зростає. Ці фактори не піддаються контролю, однак їх врахування є важливим для оцінки індивідуального ризику та планування діагностичних стратегій. Тепер розглянемо незмінні фактори ризику захворювання Альцгеймера детальніше.

Вік

Вік є найбільшим фактором ризику розвитку хвороби Альцгеймера, як і більшості видів деменції. Це означає, що людина має більшу ймовірність захворіти на хворобу Альцгеймера з віком. Після 65 років ризик розвитку хвороби Альцгеймера подвоюється приблизно кожні п'ять років. Хоча більшість людей з хворобою Альцгеймера старші за 65 років, вона також може з'явитися у молодших людей. Приблизно кожна третя людина з деменцією має хворобу Альцгеймера з раннім початком.

Стать

Жінок старше 65 років хворіє на хворобу Альцгеймера приблизно вдвічі більше, ніж чоловіків. Це головним чином тому, що жінки, як правило, живуть довше за чоловіків. Однак жінки старше 80 років все ще мають дещо вищий ризик розвитку хвороби Альцгеймера, ніж чоловіки їхнього віку. До 65 років у жінок є 20% шанс захворіти на хворобу Альцгеймера. Жінки віком у 60 років мають вдвічі більшу ймовірність розвитку хвороби Альцгеймера, ніж раку молочної залози [9]. Одна з головних причин, чому у жінок може бути вищий рівень захворювання на хворобу Альцгеймера, ніж у чоловіків, полягає в тому, що жінки в середньому живуть довше, а старший вік є значним фактором ризику розвитку цього захворювання.

Гени

Існують певні гени, які можуть передаватися у спадок від батьків і впливати на шанси людини захворіти на хворобу Альцгеймера. Існує два типи цих генів: «детерміновані» гени та гени «ризiku». Детерміновані гени майже гарантують, що люди розвинути цю хворобу, якщо проживуть достатньо довго. Люди з детермінованими генами часто розвивають симптоми хвороби Альцгеймера приблизно у віці від 40 до 50 років [10]. З 1000 людей, які хворіють на хворобу Альцгеймера, менше десяти матимуть її через детерміновані гени [11]. Гени ризику збільшують шанси людини захворіти на

хворобу Альцгеймера. Вони зустрічаються набагато частіше, ніж сімейні гени. Однак, на відміну від детермінованих генів, гени ризику не завжди спричиняють розвиток цього захворювання. Більшість із них лише незначно збільшують ризик. Люди із синдромом Дауна мають набагато вищий ризик розвитку хвороби Альцгеймера через різницю в їхніх генах. Ген, який найчастіше асоціюється з хворобою Альцгеймера, називається аполіпопротеїн АРОЕ-ε4 [12].

Сімейний анамнез

Хвороба Альцгеймера часто передається у спадок. Якщо у вас є батьки, брат/сестра або дитина з цим захворюванням, у вас більше шансів захворіти на нього. Ваш ризик зростає, якщо у кількох членів вашої родини хвороба Альцгеймера. Це може бути пов'язано з генами, факторами способу життя або поєднанням цих двох факторів. Ген АРОЕ-ε4 також відіграє тут певну роль. АРОЕ-ε4 у поєднанні з сімейним анамнезом захворювання значно підвищує ваш ризик.

Аномалії мозку

Вчені виявили аномалії мозку у людей, які, ймовірно, згодом захворіють на хворобу Альцгеймера. Одна з них - наявність крихітних скупчень білка, відомих як бляшки. Інша - скручені білкові нитки або клубки. Запалення, зменшення розміру тканин та втрата зв'язку між клітинами мозку - це інші ознаки того, що може розвинути хвороба Альцгеймера.

Модифіковані фактори ризику захворювання Альцгеймера

Окрім незмінних чинників, хвороба Альцгеймера має низку потенційно модифікованих факторів ризику, на які можна впливати шляхом зміни способу життя або впровадження профілактичних заходів на рівні суспільства. Серед них — низький рівень фізичної активності, тютюнопаління, надмірне вживання алкоголю, соціальна ізоляція, депресія, низький рівень освіти та

забруднення повітря. Згідно з останніми дослідженнями, до 45% випадків деменції можна було б попередити або відтермінувати, якби усі ці чинники були усунені або пом'якшені [13]. Багато факторів ризику також є спільними з іншими неінфекційними захворюваннями, такими як хвороби серця, рак, діабет та хронічні захворювання дихальних шляхів, що дозволяє реалізовувати інтегровані стратегії профілактики. Таким чином, підтримка активності, правильне харчування та участь у соціальній діяльності сприяють міцному здоров'ю мозку та можуть зменшити ризик розвитку захворювання Альцгеймера. Формування здорових звичок, підтримка фізичної активності, збалансоване харчування та уникнення шкідливих впливів навколишнього середовища можуть істотно зменшити ризик хвороби Альцгеймера та інших захворювань. Тепер розглянемо модифіковані фактори ризику захворювання Альцгеймера детальніше.

Фізична неактивність

Регулярна фізична активність – один із найкращих способів зменшити ризик деменції. Вона корисна для серця, кровообігу, ваги та психічного благополуччя. Дорослим рекомендується щотижня займатися помірною аеробною активністю 150 хвилин або енергійною аеробною активністю 75 хвилин.

Куріння

Куріння значно підвищує ризик розвитку деменції. Ризик інших захворювань, включаючи діабет 2 типу, інсульт, рак легенів та інші види раку також збільшується. Припинення куріння в більш пізньому віці також знижує ризик деменції.

Надмірне споживання алкоголю

Зловживання алкоголем та вживання понад 21 одиниці алкоголю на тиждень підвищує ризик деменції. Надмірне вживання алкоголю є фактором,

що сприяє розвитку понад 200 захворювань і травм. Воно має встановлений причинно-наслідковий зв'язок із різними психічними й поведінковими розладами, рядом неінфекційних хвороб, а також з підвищеним ризиком травматизму.

Забруднення повітря

Зростаюча кількість дослідницьких даних показує, що забруднення повітря підвищує ризик деменції. Необхідно пришвидшити покращення якості повітря, особливо в районах з високим рівнем забруднення повітря, оскільки це збільшує ризик виникнення захворювання Альцгеймера.

Черепно-мозкова травма

Черепно-мозкові травми (ЧМТ) є серйозною проблемою громадського здоров'я та часто виникають унаслідок дорожньо-транспортних пригод за участю автомобілів, мотоциклів і велосипедів; бойових дій; травм, отриманих під час занять контактними видами спорту; вогнепальних поранень, нападів; а також випадкових або побутових падінь. Навіть легкі або помірні повторювані травми головного мозку можуть спричинити нейрозапальні процеси, порушення обміну білків, а також загибель нейронів — ключові механізми, що лежать в основі розвитку нейродегенеративних захворювань. Пацієнти, які перенесли ЧМТ, особливо у молодому або середньому віці, мають підвищений ризик розвитку деменції у пізнішому віці, включно з хворобою Альцгеймера.

Рідкісні соціальні контакти

Недостатня соціальна активність і рідкісні соціальні контакти є значущим фактором ризику розвитку деменції, зокрема хвороби Альцгеймера. Активна соціальна взаємодія сприяє формуванню когнітивного резерву — здатності мозку компенсувати вікові або патологічні зміни за рахунок залучення альтернативних нейронних мереж. Соціальні зв'язки також стимулюють позитивну поведінку, таку як підтримка здорового способу

життя, регулярна фізична активність, дотримання режиму сну та раціонального харчування. Натомість ізоляція та обмеження соціальної взаємодії можуть призводити до зниження когнітивної стимуляції, підвищення рівня хронічного стресу та депресії, що негативно впливають на функціонування мозку.

Низький рівень освіти

Низький рівень освіти, особливо здобутої в ранньому віці, вважається одним із ключових факторів ризику розвитку деменції, включно з хворобою Альцгеймера. Встановлено, що освіта сприяє формуванню когнітивного резерву — здатності мозку ефективно протистояти нейродегенеративним процесам завдяки більш розвиненій нейронній мережі та гнучкості когнітивних функцій. Особи з вищим рівнем освіти, як правило, мають більший обсяг когнітивних ресурсів, що дозволяє їм довше зберігати функціональну незалежність навіть за наявності патологічних змін у мозку. Натомість обмежений доступ до якісної освіти або її відсутність у дитинстві знижує інтелектуальне навантаження на мозок, обмежує розвиток аналітичних навичок, мовлення, пам'яті та уваги, що може призвести до меншої стійкості нервової системи до вікових змін. Як наслідок, ризик розвитку деменції, зокрема хвороби Альцгеймера, суттєво підвищується.

Ожиріння

Особливо в середньому віці ожиріння пов'язане з підвищеним ризиком деменції. Ожиріння також пов'язане з іншими неінфекційними захворюваннями (НІЗ) і, як правило, може бути вирішене за допомогою змін способу життя, таких як дієта та фізичні вправи.

Гіпертонія

Гіпертонія (високий кров'яний тиск) у середньому віці збільшує ризик розвитку деменції, а також спричиняє інші проблеми зі здоров'ям. Ліки від

гіпертонії є єдиним відомим ефективним профілактичним засобом від деменції.

Діабет

Цукровий діабет 2 типу є добре задокументованим фактором ризику розвитку деменції, зокрема хвороби Альцгеймера. Метаболічні порушення, характерні для цього захворювання, зокрема хронічна гіперглікемія, інсулінорезистентність та запальні процеси, несприятливо впливають на стан судин головного мозку, спричиняючи порушення мозкового кровообігу та пошкодження нейронів. Таким чином, діабет 2 типу не лише шкодить серцево-судинній системі та ниркам, але також суттєво збільшує ризик когнітивного зниження, що потребує комплексного підходу до його профілактики та лікування.

Депресія

Депресія пов'язана з поширеністю деменції. Депресія є частиною продромального періоду деменції (симптом, який виникає перед симптомами, що використовуються для діагностики). Важливо контролювати та лікувати депресію, оскільки вона пов'язана зі збільшенням інвалідності, фізичними захворюваннями та гіршими наслідками для людей з деменцією.

Порушення слуху

Люди з втратою слуху мають значно підвищений ризик розвитку деменції. Використання слухових апаратів, знижує цей ризик. Оскільки втрата слуху є одним із факторів ризику, який впливає на більшість людей, її усунення може мати значний вплив на кількість людей, у яких розвивається деменція.

Невиліковна втрата зору

Невиліковна втрата зору пов'язана з підвищеним ризиком розвитку деменції. Дослідження показали, що деякі захворювання очей, такі як катаракта або діабетична ретинопатія, можуть збільшити ризик розвитку деменції. Однак це не стосується всіх захворювань очей, наприклад, глаукоми або вікової макулярної дегенерації, які, здається, не збільшують ризик. Завдяки медичному втручанню можна зменшити ризик, наприклад, дослідження показали, що ті, хто переніс операцію з видалення катаракти, мають такий самий ризик розвитку деменції, як і ті, у кого катаракта не розвивається.

Підвищений рівень ЛПНЩ

Високий рівень холестерину ліпопротеїнів низької щільності (ЛПНЩ) пов'язаний з підвищеним ризиком розвитку деменції. У мозку надлишок холестерину пов'язаний з підвищеним ризиком інсульту та збільшенням рівня амілоїду β та тау-білків, двох білків, які зазвичай пов'язані з деякими деменціями, такими як хвороба Альцгеймера.

1.4 Діагностика

1.4.1 Традиційні методи діагностики

У сфері ранньої діагностики хвороби Альцгеймера традиційні клінічні підходи відіграють ключову роль у виявленні когнітивних порушень та встановленні діагнозу. Ці методи включають оцінку медичного та сімейного анамнезу, нейропсихологічне тестування, фізичні та неврологічні обстеження, а також лабораторні аналізи з метою виключення інших причин деменції. Клінічне спостереження та оцінка когнітивного функціонування залишаються

важливою складовою діагностичного процесу. Метою цього підрозділу є висвітлення існуючих традиційних методів діагностики хвороби Альцгеймера, аналіз їхніх переваг і недоліків у сучасній клінічній практиці.

Історія хвороби

Під час первинного клінічного обстеження одним із ключових етапів є збір медичного анамнезу пацієнта, що передбачає ретельний аналіз поточного та минулого стану здоров'я. Оцінюється не лише загальний соматичний стан, але й приділяється особлива увага психіатричному анамнезу та історії когнітивних і поведінкових змін, які могли виникати впродовж життя пацієнта. Окрім того, проводиться детальний аналіз прийому лікарських засобів, що можуть впливати на когнітивні функції. Важливим компонентом обстеження є також вивчення сімейного анамнезу, зокрема наявності діагностованих випадків хвороби Альцгеймера або інших форм деменції серед близьких родичів. Встановлення спадкової схильності дозволяє підвищити точність оцінки індивідуального ризику та є одним із ключових чинників у формуванні прогностичних моделей на основі медичних даних.

Фізичний огляд та діагностичні тести

Також одним із важливих етапів медичного обстеження є фізикальне обстеження пацієнта та проведення діагностичних тестів, спрямованих на оцінку загального стану здоров'я та виявлення можливих соматичних причин когнітивних порушень. Зазвичай оцінюється харчова поведінка, нутритивний статус пацієнта, а також вживання алкоголю, що можуть мати безпосередній вплив на когнітивні функції. Обов'язковим є аналіз усього спектра медикаментозної терапії, включаючи безрецептурні засоби та харчові добавки, з метою виключення когнітивних побічних ефектів. Під час фізичного обстеження проводяться вимірювання артеріального тиску, температури тіла, частоти серцевих скорочень, а також аускультация серця та легенів. Окрім цього, виконуються базові клінічні процедури для загальної оцінки

соматичного стану. Додатково здійснюється збір зразків крові та сечі для проведення лабораторної діагностики [14]. Отримані результати дозволяють ідентифікувати стани, що можуть імітувати або супроводжувати деменцію, зокрема депресію, синдром обструктивного апное сну, делірій, побічну дію медикаментів, патології щитоподібної залози, дефіцит вітамінів, а також токсичний вплив алкоголю. На відміну від хвороби Альцгеймера, ці стани часто мають оборотний характер за умови своєчасного виявлення та належного лікування, тому їх диференціація є критично важливою на етапі попередньої діагностики.

Неврологічне обстеження

Важливим етапом є неврологічне обстеження, метою якого є виявлення ознак патологій центральної нервової системи, що можуть бути причиною когнітивних або поведінкових порушень, подібних до симптомів хвороби Альцгеймера. Під час огляду оцінюється наявність неврологічних симптомів, характерних для інсульту, хвороби Паркінсона, новоутворень головного мозку, гідроцефалії та інших станів, які можуть впливати на пам'ять, мислення та інші когнітивні функції. Неврологічне обстеження включає перевірку рефлексів, оцінку координації рухів, м'язового тону та сили м'язів, очних рухів, мовлення і сенсорної чутливості [15]. Виявлені відхилення можуть свідчити про наявність структурних або функціональних уражень мозку, що потребують подальшої диференціальної діагностики. Такий підхід є необхідним для виключення інших захворювань, клінічні прояви яких можуть імітувати деменцію, зокрема в її ранніх стадіях.

Когнітивні, функціональні та поведінкові тести

Когнітивне, функціональне та поведінкове тестування є важливою складовою діагностики хвороби Альцгеймера, оскільки дає змогу комплексно оцінити пам'ять, мислення, здатність до розв'язання простих завдань, орієнтацію в просторі й часі, а також зміни у поведінці. Такі тести

допомагають виявити порушення, що впливають на повсякденне функціонування пацієнта, зокрема на його здатність виконувати побутові завдання, приймати рішення, дотримуватися інструкцій та взаємодіяти з оточенням. Крім того, оцінюється рівень усвідомлення пацієнтом власного когнітивного дефіциту. Залежно від мети обстеження, можуть застосовуватись як короткі скринінгові інструменти, так і детальні нейропсихологічні оцінки, які зазвичай проводяться фахівцями-нейропсихологами для визначення стану виконавчих функцій, уваги, мовлення, суджень та рівня інтелектуальної активності. У стандартну практику входять такі методи, як Ascertain Dementia 8 (AD8), Functional Activities Questionnaire (FAQ), Mini-Cog, Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), Neuropsychiatric Inventory Questionnaire (NPI-Q) [16].

Скринінг на депресію та оцінка настрою

Обов'язковим елементом клінічного обстеження пацієнтів із підозрою на хворобу Альцгеймера є скринінг на депресію та оцінка емоційного стану. Це зумовлено тим, що афективні розлади, зокрема депресія, можуть спричиняти порушення пам'яті, втрату мотивації, зниження концентрації уваги та інші симптоми, які клінічно імітують деменцію або посилюють її прояви. У зв'язку з цим здійснюється оцінка загального емоційного самопочуття, щоб виявити наявність депресивних симптомів або інших порушень настрою, які можуть бути як коморбідними з когнітивними розладами, так і самостійними причинами погіршення когнітивного функціонування. Вчасне розмежування первинної деменції та псевдодеменції, зумовленої депресією, є критично важливим для правильної діагностики та вибору подальшої терапевтичної стратегії.

Аналізи спинномозкової рідини (CMP, англ. CSF)

Аналіз спинномозкової рідини (CSF) є одним із важливих інструментів у сучасній біомедичній діагностиці нейродегенеративних захворювань,

зокрема хвороби Альцгеймера. На ранніх стадіях хвороби Альцгеймера в спинномозковій рідині можуть виявлятися зміни рівнів певних біомаркерів, таких як бета-амілоїд та білок тау, які беруть участь у формуванні патологічних відкладень у головному мозку та мають високу діагностичну цінність. Ще одним перспективним біомаркером є нейрофіламент легкого ланцюга (NfL) — білок, підвищення концентрації якого спостерігається при різних нейродегенеративних станах, зокрема при хворобі Альцгеймера. На сьогодні CSF-аналізи активно застосовуються спеціалістами з деменції для підтвердження діагнозу хвороби Альцгеймера та дослідження інших форм деменції.

1.4.2 Передові методики діагностики

Останніми роками відбулося суттєве зростання інновацій у сфері діагностики хвороби Альцгеймера, що призвело до розробки передових методик, здатних забезпечити більш точне, раннє та неінвазивне виявлення патологічних змін, пов'язаних із цим захворюванням. Сучасні підходи базуються на використанні біомаркерів, нейровізуалізаційних технологій, аналізів спинномозкової рідини, тестів на основі крові, а також цифрових когнітивних інструментів і обчислювальних моделей, зокрема методів машинного навчання. Метою цього підрозділу є аналіз і систематизація новітніх діагностичних підходів до виявлення хвороби Альцгеймера, оцінка їхніх переваг, обмежень, а також вивчення їхнього потенційного впливу на практику сучасної нейродіагностики.

Аналізи крові

Аналіз крові є одним із перспективних напрямів у дослідженнях, спрямованих на розробку неінвазивних методів діагностики хвороби Альцгеймера. Наразі активно вивчається можливість виявлення стійких та вимірюваних змін у концентрації біомаркерів у периферичній крові, які

можуть корелювати зі специфічними нейродегенеративними процесами. До таких маркерів належать бета-амілоїд, тау-білок, а також індикатори нейронального ушкодження, як-от нейрофіламент легкого ланцюга (NfL) [17]. На сьогодні деякі тести на основі аналізу крові вже використовуються у спеціалізованих клініках і допомагають покращити результати клінічних випробувань. Очікується, що в майбутньому такі тести стануть невід'ємною частиною діагностичного процесу як при хворобі Альцгеймера, так і при інших формах деменції. Проте широке впровадження таких тестів в рутинну практику потребує ретельного регулювання, стандартизації та клінічної валідації, оскільки наразі дослідження в цьому напрямі ще тривають. Наявні на ринку тести можуть виявляти ознаки амілоїдної патології або нейродегенеративного ураження мозку, проте вони ще не отримали офіційного схвалення FDA. Згідно з чинними рекомендаціями, такі тести слід використовувати лише у пацієнтів із наявними когнітивними скаргами та під наглядом спеціалізованих фахівців [18]. Вони не призначені для скринінгу осіб без симптомів, і не можуть застосовуватися як самостійний метод діагностики, а лише як допоміжний інструмент у складі комплексного клінічного обстеження.

Комп'ютеризовані когнітивні тести та пристрої

Одним із перспективних напрямів сучасної нейродіагностики є розробка та впровадження комп'ютеризованих когнітивних тестів і цифрових медичних пристроїв, призначених для об'єктивної оцінки когнітивного стану пацієнтів. Такі інструменти застосовуються як у клінічній практиці, так і в межах рандомізованих клінічних досліджень, особливо в тих випадках, коли використання традиційних, тривалих за часом нейропсихологічних методик є малопрактичним або обмеженим. Комп'ютеризовані когнітивні тести забезпечують високу стандартизацію процедури оцінювання, оскільки дозволяють проводити тестування за ідентичним протоколом незалежно від клініциста. Вони дають змогу вимірювати продуктивність пацієнта під час

виконання різноманітних когнітивних і функціональних завдань, зокрема на увагу, пам'ять, швидкість обробки інформації та виконавчі функції. Такі цифрові інструменти також мають переваги у сфері масового скринінгу, дистанційної оцінки та автоматизованої обробки результатів, що особливо актуально в умовах зростання потреби у ранньому виявленні когнітивних порушень. На сьогодні Управління з контролю за продуктами і ліками США (FDA) схвалило до використання низку комп'ютеризованих когнітивних тестів і пристроїв, зокрема: Automated Neuropsychological Assessment Metrics (ANAM), Cambridge Neuropsychological Test Automated Battery (CANTAB Mobile), CognICA, Cognigram, Cognivue. Крім того, FDA схвалило також медичний пристрій Cognision, що являє собою гарнітуру з електродами, які кріпляться до шкіри голови з метою реєстрації електричної активності головного мозку, пов'язаної з когнітивним функціонуванням [19]. Інтеграція подібних цифрових інструментів у клінічну практику не лише підвищує об'єктивність когнітивної оцінки, а й створює основу для подальшого застосування методів машинного навчання для побудови прогнозних моделей на основі високоточних, кількісно вимірюваних когнітивних характеристик.

Візуалізація мозку

Структурна нейровізуалізація є важливою складовою стандартного медичного обстеження пацієнтів з підозрою на хворобу Альцгеймера. Найчастіше для цієї мети застосовуються магнітно-резонансна томографія (МРТ) або комп'ютерна томографія (КТ).

МРТ використовує радіохвилі та сильне магнітне поле для отримання детальних зображень мозку. Вони можуть показувати зменшення деяких ділянок мозку, пов'язаних з хворобою Альцгеймера. МРТ-сканування також може виключити інші стани, які можуть викликати симптоми. МРТ також проводиться перед початком прийому певних ліків від хвороби Альцгеймера та протягом усього лікування для моніторингу потенційних побічних ефектів.

В свою чергу КТ, спеціалізована рентгенівська технологія, яка створює поперечні зображення мозку.

Основною метою структурної візуалізації є виключення інших патологій головного мозку, які можуть зумовлювати когнітивні порушення, подібні до проявів захворювання Альцгеймера, але потребують принципово іншого лікування. Зокрема, за допомогою МРТ та КТ можливо виявити новоутворення, ознаки ішемічних або геморагічних інсультів, наслідки черепно-мозкової травми, а також гідроцефалію [20].

У деяких клінічних випадках застосовується також функціональна та молекулярна візуалізація, зокрема позитронно-емісійна томографія (ПЕТ) з мітками бета-амілоїду. ПЕТ-сканування дозволяє отримати зображення процесу захворювання. Під час ПЕТ-сканування в кров вводиться низькорівневий радіоактивний індикатор, щоб виявити певну особливість мозку. ПЕТ-сканування показує ділянки мозку, де поживні речовини не використовуються належним чином для отримання енергії, відомі як метаболізм. Виявлення закономірностей у ділянках низького метаболізму може допомогти розрізнити хворобу Альцгеймера та інші типи деменції. ПЕТ-візуалізація з амілоїдом може виміряти амілоїдні бляшки в мозку. Наявність підвищеного рівня амілоїдних відкладень у мозку може підтверджувати діагноз хвороби Альцгеймера, тоді як їх відсутність дозволяє розглядати інші етіології деменції. Цей тест в основному використовується в дослідженнях, але може бути використаний, якщо у людини є незвичайні або дуже ранні симптоми деменції. Також існує тау-ПЕТ-візуалізація, яка вимірює нейрофібрилярні клубки в мозку [21].

Впровадження сучасних технологій візуалізації значно розширило можливості дослідження структури та функції головного мозку, зробивши можливим раннє виявлення дегенеративних змін ще до появи клінічно виражених симптомів. У наукових дослідженнях активно вивчаються інші методи нейровізуалізації, що можуть бути використані як діагностичні інструменти та біомаркери перебігу захворювання, а також як джерела даних

для алгоритмів машинного навчання з метою підвищення точності прогнозування та оцінки прогресу хвороби Альцгеймера.

1.4.3 Роль машинного навчання у діагностиці

Машинне навчання (Machine Learning, ML) є сучасною галуззю штучного інтелекту, що стрімко розвивається і знаходить широке застосування в медицині, зокрема у сфері діагностики нейродегенеративних захворювань. Розглянемо роль машинного навчання у прогнозуванні ризику виникнення захворювання Альцгеймера. Машинне навчання вже продемонструвало ефективність [22, 23] у задачах прогнозування хвороби Альцгеймера на основі клінічних, демографічних, когнітивних та біомедичних даних. Алгоритми машинного навчання здатні обробляти великі обсяги гетерогенних даних та виявляти складні, нечіткі закономірності, які важко виявити традиційними статистичними методами. Завдяки цьому можливо здійснювати індивідуалізовану оцінку ризику, виявляти пацієнтів на доклінічній стадії захворювання та підвищувати ефективність скринінгових заходів. Окрім роботи з табличними даними, машинне навчання також використовується [24, 25] для аналізу нейровізуалізаційної інформації (МРТ, ПЕТ), де застосування методів глибокого навчання дозволяє виявляти структурні та функціональні зміни в головному мозку, характерні для хвороби Альцгеймера. Алгоритми машинного навчання можуть автоматизовано обробляти зображення, ідентифікувати зони атрофії, вимірювати об'єм мозкових структур та прогнозувати прогресування когнітивного дефіциту. Дослідження показують, що застосування машинного навчання у діагностиці хвороби Альцгеймера дозволяє досягти високих показників точності, чутливості та специфічності, а також підтримує медичних фахівців у прийнятті клінічних рішень. Використання таких моделей сприяє прискоренню діагностичного процесу, зменшенню суб'єктивного впливу людського фактора, а також дозволяє проводити масовий скринінг пацієнтів із

когнітивними скаргами. Використання машинного навчання також може пришвидшити діагностичний процес та зменшити навантаження на медичний персонал.

Отже, машинне навчання вже відіграє вагому роль у сфері діагностики та прогнозування хвороби Альцгеймера. Його застосування демонструє високу ефективність у виявленні факторів ризику, аналізі клінічних і когнітивних даних, а також у класифікації стадій захворювання. Алгоритми машинного навчання сприяють ранньому виявленню патологічних змін, що дає змогу своєчасно втручатися в перебіг хвороби. Впровадження ML-моделей у клінічну практику дозволяє підвищити точність діагностики, оптимізувати прийняття медичних рішень і покращити прогнозування когнітивних порушень, що має особливу цінність в умовах зростання поширеності нейродегенеративних захворювань. Незважаючи на численні переваги, застосування машинного навчання в клінічній практиці супроводжується певними викликами, серед яких — необхідність високоякісних, збалансованих і повних наборів даних, етичні питання обробки персональної медичної інформації, а також інтерпретованість та прозорість моделей. Проте подальше вдосконалення алгоритмів, розширення медичних баз даних та розвиток інфраструктури для впровадження цифрових рішень у сфері охорони здоров'я відкривають нові перспективи для покращення діагностики хвороби Альцгеймера на глобальному рівні.

1.5 Висновки до розділу 1

У цьому розділі було проведено ґрунтовне дослідження предметної області, пов'язаної з захворюванням Альцгеймера. Опрацьовано актуальну наукову літературу та результати сучасних досліджень щодо етіології, патофізіології, клінічних проявів і поширеності цього захворювання.

Особливу увагу приділено вивченню епідеміологічних даних, факторів ризику (як модифікованих, так і немодифікованих), а також механізмів розвитку когнітивних порушень у пацієнтів похилого віку.

У межах аналізу було розглянуто сучасні клінічні підходи до діагностики, включаючи фізикальне обстеження, нейропсихологічне тестування, лабораторні дослідження, нейровізуалізаційні методи (МРТ, ПЕТ, КТ), аналіз спинномозкової рідини, а також новітні цифрові технології, зокрема комп'ютеризовані когнітивні тести та неінвазивні біомаркери на основі аналізу крові. Вивчено потенціал зазначених методів у контексті раннього виявлення патологічних змін, характерних для хвороби Альцгеймера.

Особливу увагу було приділено огляду сучасних підходів на основі машинного навчання, які все активніше впроваджуються у медичну практику для покращення точності діагностики, індивідуалізації прогнозу та підтримки клінічного прийняття рішень. Оцінено переваги та виклики використання алгоритмів машинного навчання у сфері нейродіагностики.

Загалом, проведені дослідження предметної області дозволило сформулювати теоретичне підґрунтя для розробки моделей машинного навчання, спрямованих на прогнозування ризику виникнення хвороби Альцгеймера. Отримані знання стали важливою основою для подальших практичних етапів дослідження, зокрема для підготовки даних, побудови моделей та аналізу їх ефективності.

РОЗДІЛ 2 МАТЕМАТИЧНІ МЕТОДИ ТА МОДЕЛІ ЯК ІНСТРУМЕНТ ДІАГНОСТИКИ ЗАХВОРЮВАННЯ АЛЬЦГЕЙМЕРА

2.1 Основні поняття і термінологія машинного навчання

Машинне навчання (МН, англ. Machine Learning, ML) є одним із найбільш перспективних напрямів сучасних інформаційних технологій, який поступово трансформує традиційні підходи до аналізу, обробки та інтерпретації складних медичних даних. Цей підхід базується на розробці алгоритмів, здатних автоматично виявляти закономірності у великих масивах даних, навчатися на їх основі та формувати прогностичні моделі без необхідності явного програмування кожного кроку.

У контексті діагностики хвороби Альцгеймера машинне навчання відкриває нові можливості для раннього виявлення когнітивних порушень, аналізу ризик-факторів та індивідуалізованого прогнозування перебігу захворювання. Застосування машинного навчання у сфері нейродіагностики дозволяє опрацьовувати багатовимірні дані, що включають клінічні, когнітивні, біохімічні та візуалізаційні показники, з метою виявлення складних взаємозв'язків, які не завжди очевидні при традиційному аналізі. Такі можливості роблять машинне навчання особливо актуальним інструментом для побудови високоточних моделей прогнозування ризику виникнення хвороби Альцгеймера.

Розуміння базових концепцій машинного навчання є критично важливим для його ефективного використання в медичних дослідженнях. У цьому розділі надається загальне визначення машинного навчання, розглядається його алгоритм, типи машинного навчання, а також його зв'язок зі штучним інтелектом, нейронними мережами та глибоким навчанням.

Отримані теоретичні знання стануть основою для подальшої розробки та впровадження алгоритмів прогнозування у рамках цього дослідження.

Машинне навчання (ML) — це галузь штучного інтелекту (ШІ), що зосереджена на тому, щоб дозволити комп'ютерам та машинам імітувати спосіб навчання людей, виконувати завдання автономно та покращувати свою продуктивність і точність завдяки досвіду та доступу до більшої кількості даних. Розглянемо детальніше, як машинне навчання пов'язане з штучним інтелектом, нейронними мережами та глибоким навчанням на рис. 2.1.

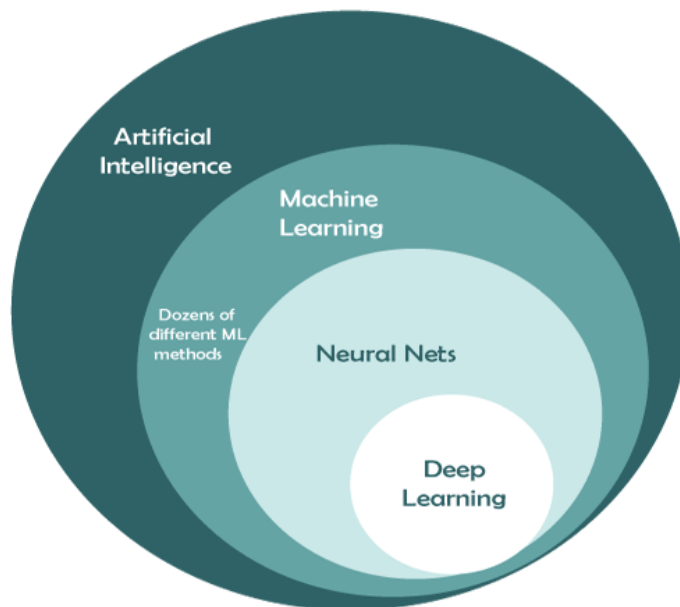


Рисунок 2.1 - Відношення між штучним інтелектом (Artificial Intelligence), машинним навчанням (Machine Learning), нейронними мережами та глибоким навчанням

Штучний інтелект (ШІ): Це найширше поняття, яке охоплює будь-яку техніку, що дозволяє машинам імітувати людську поведінку. ШІ включає широкий спектр технологій та підходів, що дозволяють комп'ютерам вирішувати проблеми, приймати рішення та розуміти мову, серед інших можливостей. Він не обмежується жодним конкретним методом чи алгоритмом і включає як системи на основі правил (де рішення приймаються на основі заздалегідь визначених правил), так і системи, які навчаються на даних.

Натомість машинне навчання (МН): МН – це підмножина ШІ, що зосереджена на ідеї, що системи можуть навчатися на даних, виявляти закономірності та приймати рішення з мінімальним втручанням людини. Воно відходить від явного програмування для конкретних завдань, натомість дозволяючи машинам коригувати та покращувати свою продуктивність на основі впливу більшої кількості даних з часом. МН включає різноманітні методи, такі як дерева рішень, регресія та нейронні мережі, серед іншого.

Нейронні мережі – це засіб машинного навчання, що дозволяє комп'ютеру освоювати виконання певних завдань шляхом аналізу навчальних прикладів, які зазвичай попередньо маркуються вручну. Нейронні мережі імітують складні функції людського мозку. Ці моделі складаються з взаємопов'язаних вузлів або нейронів, які обробляють дані, вивчають закономірності та дозволяють виконувати такі завдання, як розпізнавання образів та прийняття рішень.

В той же час глибоке навчання (ГН): ГН – це підмножина МН, яка використовує штучні нейронні мережі з багатьма шарами для моделювання складних закономірностей у даних. Глибоке навчання виявилось особливо успішним у таких завданнях, як розпізнавання зображень та мовлення, обробка природної мови, і може перевершити традиційні методи машинного навчання в обробці великих обсягів даних. Структура моделей глибокого навчання натхненна архітектурою людського мозку, але спрощена.

Машинне навчання може бути класифіковано за кількома типами, кожен з яких виконує різні завдання.

Кероване навчання (Supervised Learning) є одним з основних типів машинного навчання та характеризується використанням розмічених наборів даних, у яких кожен приклад містить як вхідні ознаки, так і відповідне цільове значення (мітку). Метою є побудова моделі, яка здатна навчатися на основі прикладів, правильно класифікувати нові дані або робити точні передбачення. У процесі навчання алгоритм поступово коригує внутрішні параметри (ваги),

щоб мінімізувати помилку між передбаченими та фактичними значеннями. Для оцінки здатності моделі до узагальнення на нових даних застосовується процедура крос-валідації, яка дозволяє уникнути проблем перенавчання (overfitting) або недонавчання (underfitting). Кероване навчання широко застосовується для вирішення практичних задач класифікації та регресії, наприклад, для діагностики захворювань та прогнозування ризиків.

Некероване навчання (Unsupervised Learning) є одним із базових типів машинного навчання, яке відрізняється від керованого тим, що використовує нерозмічені дані, тобто такі, що не містять попередньо визначених цільових міток. Основним завданням некерованого навчання є виявлення прихованих закономірностей, структур або групувань (кластерів) у даних без зовнішнього втручання чи нагляду з боку людини. Алгоритми некерованого навчання дозволяють автоматично ідентифікувати подібності, відмінності та внутрішні зв'язки між елементами вибірки, що робить цей підхід ефективним у дослідницькому аналізі даних, кластеризації користувачів, виявленні аномалій, сегментації зображень і побудові стратегій перехресного продажу. Окрім того, некероване навчання широко використовується для зменшення розмірності даних, що дозволяє підвищити ефективність моделей за рахунок усунення надлишкових ознак.

Напівкероване навчання (Semi-supervised Learning) є гібридним підходом у межах машинного навчання, що поєднує елементи керованого та некерованого навчання. Цей підхід застосовується у випадках, коли доступна лише невелика кількість розмічених даних разом із великим обсягом нерозмічених даних, що характерно для багатьох реальних сценаріїв, зокрема у медичній сфері, де маркування даних є ресурсозатратним і потребує участі кваліфікованих фахівців. Під час навчання невелика розмічена підмножина даних використовується для орієнтації моделі, яка одночасно здійснює класифікацію та виділення ознак із нерозмічених зразків. Такий підхід дозволяє значно підвищити ефективність моделі, не вимагаючи повноцінного ручного маркування великого обсягу даних, що є дорогим, повільним або

практично неможливим у деяких випадках. Напівкероване навчання особливо ефективно у тих задачах, де отримання точної розмітки є складним завданням, наприклад, при класифікації медичних зображень, обробці когнітивних або поведінкових оцінок, а також при аналізі біомедичних сигналів. Цей підхід дозволяє досягати високих результатів точності, використовуючи лише обмежену кількість маркованих прикладів у поєднанні з великою кількістю необроблених даних.

Навчання з підкріпленням (Reinforcement Learning, RL) є окремим підходом у машинному навчанні, який, на відміну від керованого навчання, не використовує розмічені приклади для безпосереднього навчання моделі. Замість цього, агент (алгоритм) взаємодіє із середовищем і навчається шляхом спроб і помилок, отримуючи зворотний зв'язок у вигляді підкріплення — позитивного або негативного сигналу (нагороди або штрафу), залежно від обраної дії. Метою навчання з підкріпленням є розробка оптимальної політики дій, яка дозволяє максимізувати сумарну винагороду впродовж певного періоду часу. У цьому контексті RL-агент поступово накопичує досвід і формує стратегію прийняття рішень, яка базується на послідовності дій та відповідних результатах. Таким чином, навчання з підкріпленням є потужним інструментом у випадках, коли модель повинна самостійно вивчати стратегічну поведінку в умовах часткової або повної невизначеності, що відкриває перспективи його застосування і в медичних задачах, наприклад, для оптимізації схем лікування чи адаптивної когнітивної реабілітації.

Згідно з підходом, запропонованим Каліфорнійським університетом у Берклі, процес навчання алгоритму машинного навчання із учителем зазвичай складається з трьох основних компонентів.

1. Процес прийняття рішень.

На першому етапі алгоритм виконує певну послідовність обчислень та операцій, що базуються на вхідних даних — як розмічених, так і нерозмічених. Метою є формування попереднього припущення (гіпотези) щодо структури

закономірностей у даних. На основі вхідних даних алгоритм надасть оцінку щодо закономірності в даних.

2. Функція помилки.

Цей компонент відповідає за кількісну оцінку точності передбачення, що здійснюється моделлю. За наявності еталонних (істинних) відповідей функція помилки дозволяє порівняти прогноз із фактичним результатом та обчислити рівень похибки. Таким чином, вона є критерієм якості моделі на кожному кроці навчання.

3. Процес оптимізації моделі.

Після оцінки похибки, якщо модель може точніше відтворювати дані з навчального набору, відбувається коригування ваг з метою зменшення розбіжності між відомими прикладами та прогнозом моделі. Алгоритм багаторазово повторює цей ітеративний процес «оцінювання та оптимізації», самостійно оновлюючи ваги до досягнення заданого порогу точності.

У сукупності ці три компоненти формують основу механізму навчання, що забезпечує здатність моделей машинного навчання поступово покращувати свої прогнози шляхом адаптації до структури вхідних даних. Такий підхід дозволяє будувати моделі, здатні до узагальнення й ефективного застосування на нових, раніше невідомих прикладах.

2.2 Математичні основи машинного навчання

Машинне навчання спирається на ґрунтовні математичні концепції, які формують теоретичну основу алгоритмів, моделей та методів, що використовуються для аналізу даних і прийняття рішень. Саме математичний апарат дозволяє формалізувати процес навчання, обробки інформації, побудови прогнозів та оптимізації моделей. Розуміння математичних

принципів є необхідною передумовою для ефективної реалізації алгоритмів машинного навчання. У цьому розділі буде розглянуто ключові математичні дисципліни, що лежать в основі машинного навчання, зокрема: лінійну алгебру, математичний аналіз, теорію ймовірностей, математичну статистику, теорію оптимізації, а також алгоритми та структури даних. Їх вивчення дає змогу глибше осмислити внутрішні механізми роботи моделей, а також забезпечити коректність і точність побудови систем прогнозування на основі даних.

1. Математичний аналіз.

Математичний аналіз відіграє важливу роль у формалізації та реалізації процесів навчання в алгоритмах машинного навчання. Його методи забезпечують теоретичне підґрунтя для задач оптимізації, які є центральними у побудові й налаштуванні моделей. Зокрема, математичний аналіз дозволяє формалізувати цільові функції, що описують помилку або втрати моделі, а також знайти екстремальні значення цих функцій — тобто точки, у яких модель демонструє найкращі показники. Ключові поняття математичного аналізу, такі як похідна, градієнт, частинна похідна, дозволяють здійснювати пошук напрямку найшвидшого зменшення помилки, що є основою для реалізації градієнтного спуску — одного з найпоширеніших алгоритмів оптимізації в машинному навчанні. Цей підхід дає змогу ітеративно оновлювати параметри моделі, зменшуючи різницю між прогнозованими та реальними значеннями.

2. Лінійна алгебра.

Лінійна алгебра є однією з ключових математичних дисциплін, яка становить теоретичну основу більшості алгоритмів машинного навчання. Її поняття та методи активно застосовуються як у процесі побудови моделей, так і під час обробки, подання та аналізу даних. Основні об'єкти лінійної алгебри — вектори, матриці, тензори, лінійні перетворення — використовуються для представлення ознак, структур даних та операцій над ними. У контексті машинного навчання вектори виступають у ролі репрезентації окремих

об'єктів, наприклад пацієнтів, з відповідними числовими характеристиками (ознаками), тоді як матриці відображають повні набори даних або ваги моделей. Операції, такі як додавання векторів, скалярний добуток, множення матриць, розклад матриць, обчислення власних значень та власних векторів, мають прикладне значення в алгоритмах класифікації, зниження розмірності, оптимізації та обробки зображень.

3. Теорія ймовірності.

Теорія ймовірностей є фундаментальною складовою машинного навчання, оскільки вона дозволяє формалізувати та кількісно описувати невизначеність, випадковість і ризик, які властиві багатьом реальним процесам, зокрема в медичній діагностиці. За допомогою ймовірнісного апарату моделюється поведінка складних систем при наявності неповних або нечітких даних, а також визначаються ймовірності належності об'єктів до певних класів. Ключові поняття, такі як умовна ймовірність та ймовірнісні розподіли, використовуються в алгоритмах класифікації, оцінки ризиків і побудови байєсівських моделей.

4. Математична статистика.

Математична статистика забезпечує інструменти для обробки, узагальнення та інтерпретації емпіричних даних, що є критично важливим при навчанні моделей машинного навчання. Статистичні методи дозволяють аналізувати структуру даних, оцінювати параметри розподілів, здійснювати перевірку статистичних гіпотез, а також визначати достовірність і значущість результатів.

5. Теорія оптимізації.

Теорія оптимізації є невід'ємною складовою машинного навчання, оскільки вона забезпечує методологічну основу для пошуку оптимальних рішень. У загальному випадку, оптимізація полягає у знаходженні такого набору параметрів моделі, який мінімізує (або максимізує) задану цільову функцію, що найчастіше відображає рівень похибки або втрат під час навчання алгоритму. У контексті машинного навчання оптимізаційні задачі виникають

на всіх етапах — від підбору вагових коефіцієнтів у нейронних мережах, до налаштування гіперпараметрів у моделях. Застосування алгоритмів градієнтного спуску та методів стохастичної оптимізації дозволяє ефективно вирішувати складні задачі навчання навіть у багатовимірному просторі параметрів.

6. Алгоритми та структури даних.

Алгоритми та структури даних є критично важливими компонентами машинного навчання, що забезпечують ефективність обчислень, збереження, організацію та обробку великих обсягів даних. Алгоритми визначають послідовність дій, необхідних для виконання завдання — від попередньої обробки даних до навчання моделі, тоді як структури даних забезпечують оптимальне представлення і доступ до інформації в оперативній пам'яті. У задачах машинного навчання використовуються численні алгоритмічні підходи — сортування, пошук, перебір, динамічне програмування, жадібні алгоритми тощо, які забезпечують побудову, навчання та оптимізацію моделей. Ефективність реалізації таких алгоритмів значною мірою залежить від правильного вибору структур даних, наприклад, масивів, списків, стеків, дерев, графів, хеш-таблиць.

Розуміння вищенаведених математичних основ забезпечує глибше усвідомлення принципів роботи алгоритмів машинного навчання, а також дозволяє обґрунтовано підходити до вибору моделей, їх навчання та інтерпретації результатів. Такі знання є надзвичайно важливими у сфері медичних досліджень, де точність, узагальнюваність і достовірність рішень відіграють критичну роль. У подальшій частині цього розділу буде розглянуто алгоритми машинного навчання, які застосовуються для прогнозування ризику виникнення хвороби Альцгеймера, а також проаналізовано їхні особливості, переваги, обмеження та ефективність у контексті завдань медичної діагностики.

2.3 Алгоритми машинного навчання

Існує велика кількість алгоритмів машинного навчання, кожен з яких має свої теоретичні основи, переваги, обмеження та сфери ефективного застосування. Вибір алгоритму значною мірою залежить від характеристик даних, цілей моделювання та вимог до інтерпретованості або точності результатів. У контексті прогнозування ризику виникнення хвороби Альцгеймера особливу увагу слід приділяти алгоритмам, здатним ефективно працювати з медичними, клінічними та демографічними даними, враховуючи багатофакторну природу захворювання. В даній дипломній роботі буде детально розглянуто такі п'ять алгоритмів машинного навчання: Decision Tree, Random Forest, XGBoost, Support Vector Machine, Stacking Ensemble.

2.3.1 Decision Tree

Визначення Decision Tree

Дерева рішень – це широко використовуваний алгоритм машинного навчання, який можна застосовувати як для задач класифікації, так і для задач регресії. Ці моделі працюють шляхом розділення даних на підмножини на основі ознак, і це розділення називається прийняттям рішень, де кожен кінцевий вузол повідомляє нам про прогноз. Це розділення створює деревоподібну структуру, зображену на рис. 2.2.

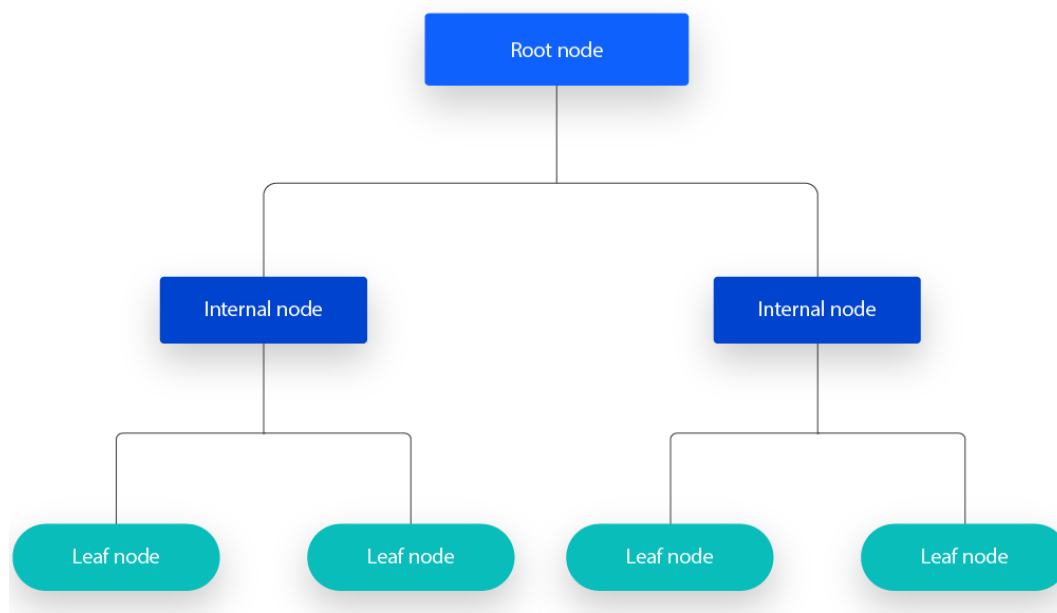


Рисунок 2.2 - Структура дерева рішень [26]

Дерево рішень – це графічне представлення різних варіантів вирішення проблеми, яке показує, як різні фактори пов'язані між собою. Воно має ієрархічну деревоподібну структуру, яка починається з одного кореневого вузла на вершині, що далі поділяється на різні можливі результати, де:

- 1) **кореневий вузол (Root Node)** – це початкова точка, яка представляє весь набір даних;
- 2) **гілки (Branches)** – це лінії, що з'єднують вузли. Вони показують перехід від одного рішення до іншого;
- 3) **внутрішні вузли (Internal Nodes)** – це точки, де приймаються рішення на основі вхідних ознак;
- 4) **листові вузли (Leaf Nodes)** – це кінцеві вузли в кінці гілок, які представляють остаточні результати або прогнози.

Метою використання дерева рішень є побудова навчальної моделі, здатної прогнозувати значення або клас цільової змінної шляхом виявлення набору простих логічних правил, що виводяться з навчальних даних. Цей підхід дозволяє моделі здійснювати послідовні рішення на основі значень

вхідних змінних. У процесі класифікації нового об'єкта дерево рішень починає з кореневого вузла та порівнює значення відповідного атрибута об'єкта з умовою, що визначена в цьому вузлі. Залежно від результату порівняння, алгоритм переходить до відповідної гілки дерева та переходить до наступного вузла. Цей процес триває до досягнення листового вузла, який містить остаточне рішення — прогнозований клас або значення цільової змінної.

Математичне обґрунтування

Ключовим елементом роботи дерева рішень у задачах класифікації є процес поділу даних (splitting), який полягає в послідовному розбитті вибірки на підмножини, що мають якомога більшу однорідність у межах класів. Для визначення доцільності поділу даних на кожному етапі використовується поняття ентропії, яке є кількісною мірою невизначеності або неупорядкованості в даних.

Ентропія оцінює рівень змішаності класів у наборі даних. Наприклад, якщо в підмножині міститься порівну об'єктів двох різних класів, то така підмножина має високу ентропію, що вказує на високий рівень невизначеності. Натомість, якщо усі об'єкти належать до одного класу, ентропія буде мінімальною або нульовою, що свідчить про повну впорядкованість.

Наведемо формулу обрахунку ентропії:

$$H(S) = - \sum_{i=1}^c p_i \log_2(p_i),$$

де p_i — це ймовірність появи об'єкта, що належить до i -го класу, c — кількість класів, $H(S)$ — це ентропія, або наскільки «змішаний» набір даних.

Таким чином, ентропія є критерієм, який використовується для вибору ознаки та порогу поділу в кожному вузлі дерева: чим нижча ентропія після поділу, тим кращим вважається цей поділ.

Далі в алгоритмах дерева рішень використовується інформаційна вигода (Information Gain) для визначення найкращої ознаки, за якою слід виконати поділ даних на кожному етапі побудови дерева. Після обчислення ентропії множини даних, інформаційна вигода дозволяє оцінити, наскільки ефективно конкретна ознака зменшує невизначеність у даних в результаті поділу.

Інформаційна вигода відображає зменшення ентропії після поділу даних за певною ознакою. Чим сильніше зменшується ентропія, тим кращим вважається поділ. Наведемо формулу обрахунку інформаційної вигоди:

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

де $IG(S, A)$ — інформаційна вигода при використанні ознаки A , $H(S)$ — ентропія початкової множини S , S_v — підмножина S , що містить об'єкти з конкретним значенням v ознаки A після поділу, $|S_v|$ — кількість елементів у підмножині S_v після поділу, $|S|$ — загальна кількість елементів у множині S .

Розглянемо порівняння низької інформаційної вигоди (low information gain) і високої інформаційної вигоди (high information gain) на рис.2.3.

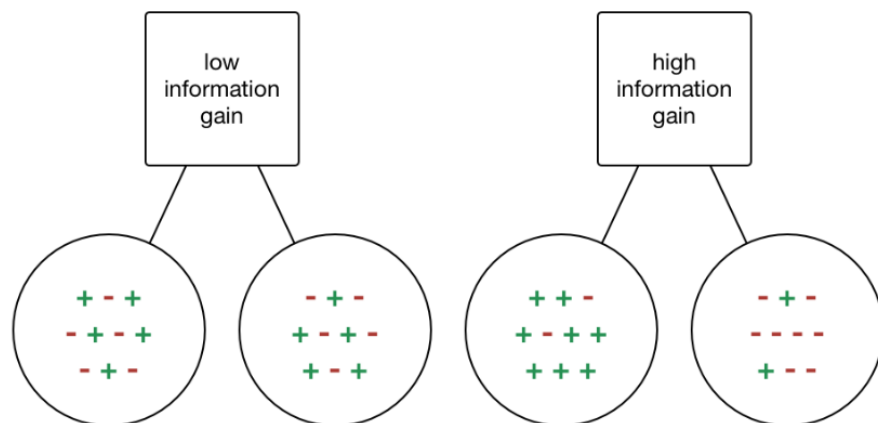


Рисунок 2.3 – Порівняння низької інформаційної вигоди (low information gain) і високої інформаційної вигоди (high information gain) [27]

Алгоритм перебирає всі можливі ознаки і для кожної з них обчислює інформаційну вигоду. Той поділ, який забезпечує найбільшу вигоду, тобто найбільше зменшення ентропії, обирається як оптимальний. У такий спосіб дерево намагається забезпечити найменшу можливу змішаність класів у результаті поділу, що сприяє підвищенню точності моделі.

Індекс Джині (Gini Impurity) є альтернативним до ентропії критерієм, який широко використовується в алгоритмах дерева рішень для оцінки якості поділу даних. Він забезпечує кількісну оцінку ступеня неоднорідності множини даних. Індекс Джині дозволяє визначити, наскільки ймовірно, що випадково вибраний елемент з підмножини буде неправильно класифікований, якщо класифікацію здійснювати за розподілом класів у цій підмножині.

Наведемо формулу обрахунку індексу Джині:

$$Gini(S) = 1 - \sum_{i=1}^c p_i^2$$

де p_i — ймовірність того, що випадково обраний об'єкт належить до класу i . Значення Gini-індексу коливається від 0 (ідеальна чистота) до 1 (максимальна неоднорідність).

Як і у випадку з інформаційною вигодою, при побудові дерева рішень розглядаються всі можливі ознаки, і обирається така, яка після поділу забезпечує найменше значення Gini-індексу для утворених підмножин. Порівняно з ентропією, індекс Джині часто обчислюється швидше, що робить його популярним вибором у практичних реалізаціях.

Переваги Decision Tree

Дерева рішень мають низку суттєвих переваг, які обумовлюють їх широку застосовність у задачах машинного навчання. Насамперед, вони

відзначаються простотою реалізації та високою інтерпретованістю. Модель можна легко представити у вигляді деревоподібної структури або блок-схеми, що дозволяє наочно відстежувати логіку прийняття рішень. Крім того, дерева рішень є універсальними алгоритмами, які можуть ефективно застосовуватись як до задач класифікації, так і до задач регресії. Важливою перевагою є також те, що вони не потребують попереднього масштабування або нормалізації ознак, що спрощує етап підготовки даних. Окрім того, дерева рішень здатні моделювати складні, нелінійні взаємозв'язки між вхідними ознаками та цільовою змінною, що робить їх ефективними у випадках, коли залежності між параметрами не є лінійними.

Недоліки Decision Tree

Попри численні переваги, дерева рішень мають і низку обмежень, які слід враховувати при їх застосуванні. Однією з основних проблем є схильність до перенавчання (overfitting), коли модель надмірно точно відображає структуру навчальних даних, включаючи випадкові шумові значення, що призводить до погіршення її узагальнювальної здатності на нових даних. Ще одним недоліком є нестабільність моделі, тобто чутливість до незначних змін у вхідних даних, які можуть спричинити суттєві відмінності у структурі дерева та, відповідно, у прогнозах. Крім того, дерева рішень можуть виявляти систематичну упередженість на користь ознак з великою кількістю категорій, що може призвести до надмірної концентрації моделі на таких ознаках, ігноруючи інші, потенційно більш значущі предиктори, що в кінцевому результаті знижує точність прогнозування.

Придатність для прогнозування ризику виникнення захворювання Альцгеймера

Модель дерева рішень (Decision Tree) демонструє високу ефективність у медичних застосуваннях, зокрема у задачах прогнозування ризику виникнення захворювання Альцгеймера. Основними перевагами цього методу є

інтерпретованість, прозорість процесу прийняття рішень та відсутність необхідності попереднього масштабування або нормалізації даних. У контексті прогнозування нейродегенеративних захворювань дерево рішень дозволяє побудувати логічну структуру на основі клінічних, демографічних, когнітивних та функціональних ознак, що значно спрощує інтерпретацію моделі фахівцями у сфері охорони здоров'я.

Крім того, Decision Tree добре працює з невеликими та неповними вибірками, що особливо важливо в медичних дослідженнях, де доступ до повних і збалансованих даних може бути обмеженим. Завдяки здатності моделювати нелінійні залежності між ознаками, дерево рішень ефективно виявляє взаємозв'язки між факторами ризику та ймовірністю виникнення хвороби Альцгеймера. У поєднанні з належним налаштуванням глибини дерева та механізмами обрізання (pruning), цей метод здатен забезпечити високу точність класифікації при збереженні узагальнюючої здатності, що робить його цінним інструментом для підтримки клінічних рішень у сфері когнітивної діагностики.

2.3.2 Random Forest

Визначення Random Forest

Random Forest (RF) — це один із потужних алгоритмів машинного навчання, що застосовується у задачах керованого навчання, тобто для побудови моделей на основі розмічених даних з метою виконання класифікації або регресії. Основу методу становить ансамблевий підхід, побудований на великій кількості дерев рішень, які поєднуються в єдину модель для підвищення точності та стійкості до перенавчання. Структура моделі Random Forest зображена на рис. 2.4.

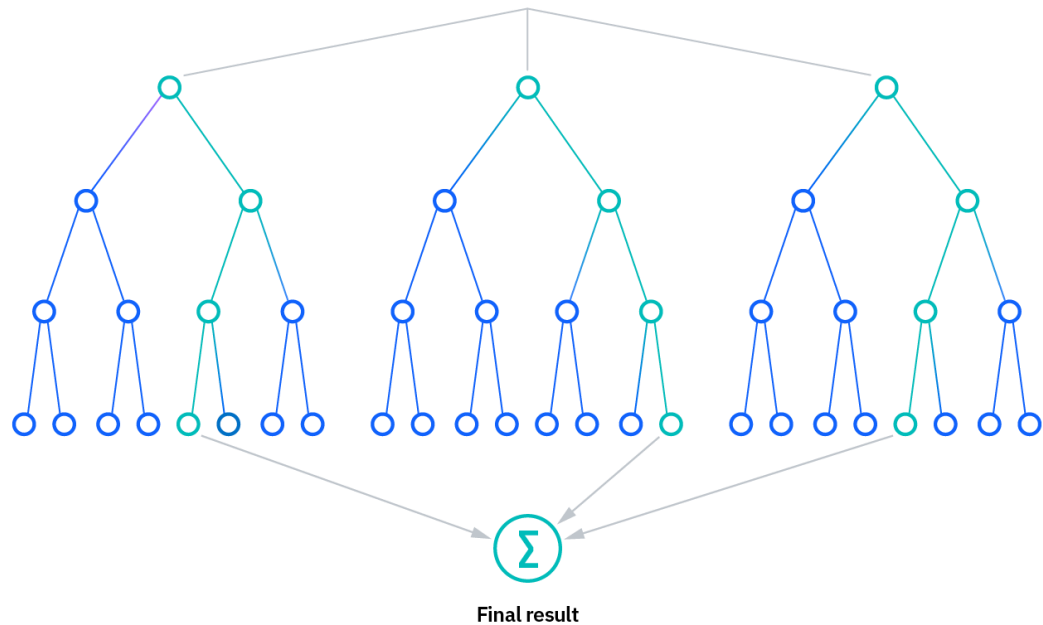


Рисунок 2.4 - Структура Random Forest [28]

Алгоритм Random Forest реалізує принцип берінгу (bagging) — методу, згідно з яким кожне дерево навчається на випадковій підмножині навчальної вибірки, а також використовує випадкову вибірку ознак під час побудови кожного дерева. Така стратегія сприяє формуванню некорельованих дерев, що знижує загальну дисперсію моделі та підвищує її узагальнюючу здатність. Остаточне рішення моделі формується на основі голосування за результати, отримані від окремих дерев класифікатора.

Random Forest поєднує в собі високу точність, стійкість до шуму у даних, а також низьку ймовірність перенавчання у порівнянні з окремими деревами рішень. Ці властивості роблять його ефективним інструментом для задач медичної діагностики, зокрема при аналізі складних взаємозв'язків між клінічними, демографічними та поведінковими факторами, що мають значення для прогнозування ризику виникнення хвороби Альцгеймера.

Кожне окреме дерево в ансамблі Random Forest навчається на випадковій підмножині навчального набору, бутстреп-вибірці (bootstrap sample). Близько однієї третини прикладів, які не потрапили до цієї вибірки,

утворюють out-of-bag вибірку, яка використовується для оцінювання якості моделі без необхідності окремої валідаційної множини.

Крім того, під час побудови кожного дерева реалізується випадковий вибір підмножини ознак (feature bagging), що сприяє зниженню кореляції між деревами та, відповідно, підвищенню узагальнюючої здатності ансамблю. Залежно від типу задачі, об'єднання результатів окремих дерев здійснюється по-різному: у випадку регресії — шляхом усереднення прогнозів, у випадку класифікації — шляхом більшості голосів.

Використання out-of-bag вибірки як механізму вбудованої крос-валідації дозволяє оцінити ефективність моделі без додаткового розподілу на навчальну та тестову вибірки, що особливо цінне у медичних дослідженнях, де обсяг даних часто є обмеженим.

Математичне обґрунтування

Алгоритм Random Forest є ансамблевим методом, що ґрунтується на поєднанні множини дерев рішень, кожне з яких навчається на випадковій підмножині даних.

Позначимо навчальний набір даних як

$$X = \{(x_i, y_i)\}_{i=1}^n ,$$

де x_i - вектор ознак, y_i - відповідна мітка класу.

Прогноз кожного дерева рішень у ансамблі можна розглядати як функцію:

$$f_t(x, \theta_t) ,$$

де θ_t - набір параметрів, який визначає структуру t-го дерева у лісі, $f_t(x, \theta_t)$ — прогноз класу для вхідного вектора x .

Прогноз окремого дерева рішень є бінарною функцією, що повертає прогнозовану мітку класу для заданого вхідного значення x . У задачах класифікації ця мітка може набувати значення 0 або 1. У методі випадкового лісу прогнози численних дерев рішень об'єднуються за допомогою схеми голосування більшості. Остаточне передбачення $\hat{y}(x)$ для конкретного вхідного значення x визначається як той клас, що отримав більшість голосів серед усіх дерев ансамблю.

Оскільки ансамбль дерев рішень навчається з метою мінімізації загальної похибки прогнозування, то цільова функція навчання моделі Random Forest має вигляд:

$$\min_{\theta_1, \dots, \theta_m} \sum_{i=1}^n L(y_i, \hat{y}(x_i)),$$

де $L(y_i, \hat{y}(x_i))$ – функція втрат, що вимірює відхилення між справжнім класом y_i та прогнозованим класом $\hat{y}(x_i)$, а m – кількість дерев у ансамблі.

Переваги Random Forest

Однією з ключових переваг алгоритму Random Forest є зменшення ризику перенавчання (overfitting), характерного для окремих дерев рішень, які можуть надмірно точно відтворювати тренувальні дані. У випадку використання ансамблю дерев, випадковість та агрегування прогнозів знижують кореляцію між окремими деревами, що, у свою чергу, зменшує дисперсію та підвищує узагальнюючу здатність моделі. Крім того, алгоритм відзначається високою гнучкістю, оскільки ефективно працює як у задачах класифікації, так і регресії. Завдяки механізму випадкового вибору ознак (feature bagging), Random Forest добре справляється з неповними даними, зберігаючи високу точність навіть за наявності пропущених значень.

Важливою характеристикою моделі є також, що випадковий ліс дозволяє легко оцінити важливість ознак і їх внесок у модель.

Недоліки Random Forest

Попри численні переваги, алгоритм Random Forest має й певні обмеження, які слід враховувати під час його застосування. Насамперед, обчислювальна складність моделі зростає пропорційно до кількості дерев у лісі, що може призводити до значного споживання ресурсів пам'яті та збільшення часу обробки, особливо при роботі з великими наборами даних. Кожне дерево потребує окремого обчислення, а об'єднання результатів усіх дерев може суттєво впливати на швидкодію системи. Крім того, у порівнянні з одиничним деревом рішень, інтерпретація результатів Random Forest є менш прозорою, що ускладнює пояснення моделі в критично важливих галузях, таких як медицина. Ще одним потенційним недоліком є те, що окремі дерева всередині лісу можуть бути схильними до перенавчання, хоча сам Random Forest завдяки використанню бутстреп-вибірок і випадкової вибірки ознак зазвичай успішно запобігає цьому ефекту. Отже, хоча Random Forest є надійним і потужним методом, його застосування вимагає зваженого підходу до вибору параметрів та обсягу вхідних даних, з урахуванням наявних обчислювальних ресурсів і вимог до інтерпретованості моделі.

Придатність для прогнозування ризику виникнення захворювання Альцгеймера

Модель Random Forest є одним з ефективних та широко застосовуваних методів у задачах медичної класифікації, зокрема для прогнозування ризику виникнення хвороби Альцгеймера. Її висока стійкість до перенавчання, здатність працювати з високорозмірними, неповними або гетерогенними даними, а також вбудовані механізми оцінки важливості ознак роблять цей алгоритм особливо корисним у клінічних дослідженнях, де структура даних є складною та багатofакторною.

У медичній практиці, де важливо враховувати як демографічні, так і клінічні та когнітивні показники пацієнтів, Random Forest дозволяє моделювати нелінійні залежності між численними вхідними змінними та цільовим прогнозом. Завдяки ансамблевій структурі, цей алгоритм поєднує рішення багатьох дерев рішень, навчених на випадкових підмножинах даних та ознак, що забезпечує високу узагальнюючу здатність та зменшує кореляцію між окремими моделями. Крім того, механізм out-of-bag оцінювання дозволяє отримати надійні оцінки точності моделі без потреби у додатковій валідаційній вибірці, що є особливо важливим у випадках обмеженого обсягу доступних даних.

Random Forest також надає можливість кількісної оцінки значущості кожної ознаки, що дозволяє дослідникам та лікарям ідентифікувати найбільш впливові чинники ризику, пов'язані з розвитком хвороби Альцгеймера. Це не лише підвищує прогностичну цінність моделі, а й сприяє кращому розумінню патофізіологічних механізмів захворювання. Таким чином, модель Random Forest є доцільним вибором для реалізації задач прогнозування в галузі нейродегенеративних захворювань і має високий потенціал для використання в системах підтримки прийняття медичних рішень.

2.3.3 XGBoost

Визначення XGBoost

XGBoost (Extreme Gradient Boosting) — це ефективна реалізація методу градієнтного бустингу, який належить до класу ансамблевих алгоритмів машинного навчання. Основна ідея методу полягає в послідовному поєднанні великої кількості слабких моделей, зазвичай дерев рішень з обмеженою глибиною, з метою побудови сильної узагальнюючої моделі. Кожна нова модель додається до ансамблю таким чином, щоб скоригувати похибки попередніх моделей, що дозволяє поступово зменшувати загальну помилку прогнозування.

На відміну від методів бутстреп-ансамблювання (наприклад, Random Forest), де дерева будуються незалежно та паралельно, у XGBoost реалізується послідовне навчання, тобто кожне нове дерево оптимізується з урахуванням похибок попереднього. Механізм навчання базується на градієнтному спуску, який забезпечує мінімізацію функції втрат за допомогою обчислення похідних (градієнтів) цієї функції щодо параметрів моделі. Таким чином, на кожному етапі навчання алгоритм спрямовується у напрямку найшвидшого зменшення помилки.

Навчання починається з рівномірного розподілу ваг серед усіх прикладів навчальної вибірки. Після побудови першої моделі, ваги прикладів оновлюються відповідно до точності передбачення, тобто краще класифіковані об'єкти отримують менші ваги, а ті, що були класифіковані з помилкою — більші. Це дозволяє наступній моделі зосередитися на «важких» прикладах, які були неправильно класифіковані раніше. Такий підхід забезпечує фокусування ансамблю на найбільш інформативних об'єктах.

XGBoost також вирізняється високою обчислювальною ефективністю, можливістю регуляризації, обробкою пропущених значень, підтримкою паралельного обчислення та масштабованістю, що забезпечує його успішне застосування у практичних задачах з великим обсягом даних. У зв'язку з цим, XGBoost є одним із провідних інструментів для побудови високоточних моделей прогнозування, зокрема в задачах медичної діагностики, де критично важливо досягти високої точності та узагальнюваності результатів.

Математичне обґрунтування

Алгоритм XGBoost реалізує підхід керованого навчання з використанням методу градієнтного бустингу для побудови ансамблю дерев рішень. Алгоритм XGBoost є ітеративним ансамблевим методом підсилення, який базується на додаванні нових дерев рішень з метою зменшення похибки моделі. Початкове прогнозування зазвичай встановлюється рівним нулю,

після чого на кожному кроці до моделі додається нове дерево, яке прагне зменшити залишкову похибку попередньої моделі.

Математично модель представлена як сума прогнозів усіх дерев:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i)$$

де $\hat{y}_i$ — підсумкове прогнозне значення для i -го прикладу, K — кількість дерев у ансамблі, $f_k(x_i)$ — прогноз k -го дерева для прикладу x_i .

Цільова функція XGBoost складається з двох компонентів: функції втрат (loss function) та члена регуляризації (regularization term). Функція втрат оцінює, наскільки добре модель наближає істинні значення, а регуляризація контролює складність дерев, запобігаючи перенавчанню. Загальний вигляд цільової функції:

$$obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

де $l(y_i, \hat{y}_i)$ — функція втрат між істинним значенням y_i та прогнозом $\hat{y}_i$, $\Omega(f_k)$ — член регуляризації, який карає занадто складні дерева.

Тепер, замість того, щоб підігнати модель одразу, ми оптимізуємо її ітеративно. Ми починаємо з початкового прогнозу. $\hat{y}_i^{(0)} = 0$ на кожному кроці ми додаємо нове дерево для покращення моделі. Оновлені прогнози після додавання t -го дерева можна записати так:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

де $\hat{y}_i^{(t-1)}$ — прогноз з попередньої ітерації, $f_t(x_i)$ — це прогноз t -го дерева для i -ї точки даних.

Член регуляризації $\Omega(f_t)$ спрощує складні дерева, обмежуючи кількість листків у дереві та розмір листка. Це визначається як:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2$$

де T — кількість листків у дереві, γ — параметр регуляризації, що контролює складність дерева, λ — коефіцієнт, який штрафує великі ваги ω_j листків.

Під час побудови дерева важливим кроком є обчислення приросту інформації (information gain), який визначає доцільність кожного розбиття вузла. Інформаційний приріст для розділення обчислюється так:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

де G_L, G_R — суми градієнтів для лівого та правого підвузлів відповідно, H_L, H_R — суми Гессіанів (других похідних) для лівого та правого підвузлів.

Таким чином, алгоритм XGBoost на кожному кроці обирає таке розбиття вузла, яке забезпечує максимальний приріст інформації. Це дозволяє ефективно зменшувати похибку моделі та досягати високої точності прогнозування. Візуально алгоритм XGBoost продемонстровано на рис. 2.5.

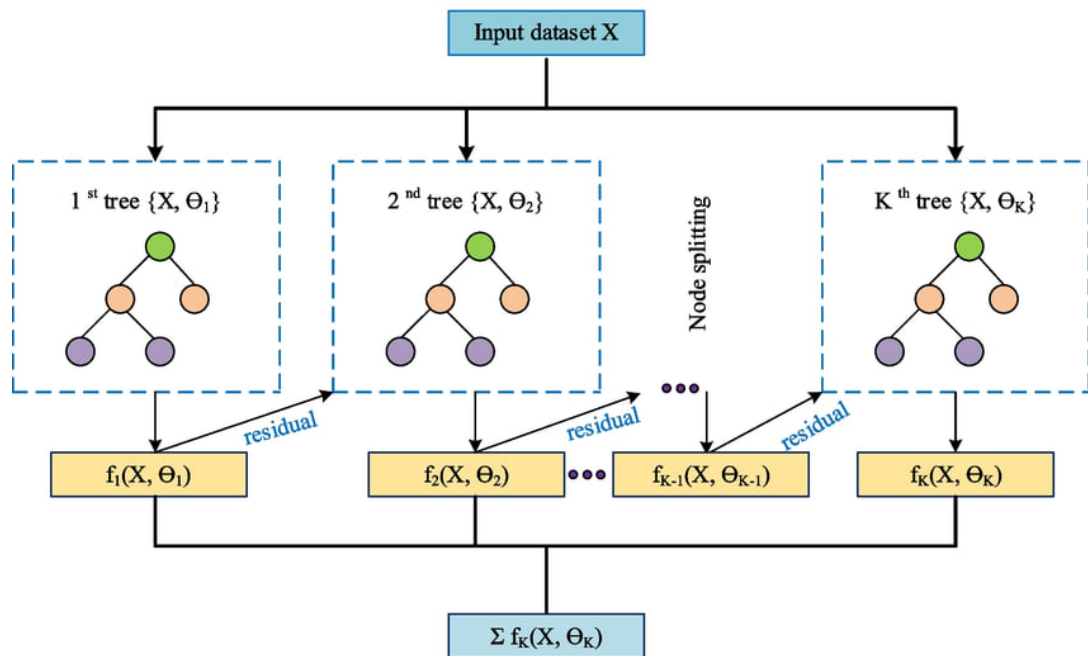


Рисунок 2.5 - Алгоритм XGBoost

Переваги XGBoost

Алгоритм XGBoost вирізняється високою продуктивністю, ефективністю та широким спектром застосування у задачах машинного навчання. Завдяки оптимізованому механізму побудови дерев та реалізації градієнтного бустингу з регуляризацією, XGBoost часто демонструє вищу точність прогнозування порівняно з іншими алгоритмами, особливо на наборах даних зі змішаними типами ознак — числовими, категоріальними та порядковими. Однією з визначальних переваг є висока швидкодія, що досягається за рахунок можливості паралельної обробки за допомогою багатоядерних процесорів. Крім того, XGBoost має вбудовану здатність до обробки пропущених значень, що знижує потребу в попередньому очищенні даних. Ще однією перевагою є вбудована регуляризація, яка дозволяє контролювати складність моделі та зменшує ймовірність перенавчання, що забезпечує високу узагальнюючу здатність моделі на нових, раніше невідомих даних. Завдяки поєднанню точності, швидкості та адаптивності, XGBoost є одним із найбільш надійних алгоритмів для вирішення складних прикладних задач, зокрема у сфері медичного прогнозування.

Недоліки XGBoost

Попри численні переваги, алгоритм XGBoost має і низку обмежень, які слід враховувати при його застосуванні в практичних задачах. Зокрема, модель може бути схильна до перенавчання, особливо у випадках, коли гіперпараметри налаштовані неналежним чином. Хоча XGBoost включає механізми регуляризації, ефективність моделі значною мірою залежить від ретельного підбору параметрів, таких як швидкість навчання, глибина дерев та кількість ітерацій. Процес налаштування є нетривіальним і часто потребує застосування методів автоматизованого пошуку, зокрема ґрид-пошуку (grid search) або випадкового пошуку (randomized search), що може бути ресурсозатратним. Крім того, незважаючи на те, що окремі дерева, які входять до складу ансамблю, можна інтерпретувати, загальна модель XGBoost має низьку інтерпретованість, що ускладнює пояснення логіки прийняття рішень у порівнянні з лінійними моделями або окремими деревами рішень. Цей аспект є особливо важливим у медичних застосуваннях, де прозорість моделі має критичне значення для обґрунтування діагностичних рішень.

Придатність для прогнозування ризику виникнення захворювання Альцгеймера

Модель XGBoost (Extreme Gradient Boosting) є однією з найефективніших сучасних реалізацій градієнтного бустингу та широко використовується для вирішення задач класифікації у медичній сфері. Алгоритм поєднує в собі високу точність, стійкість до шуму та здатність працювати з великими і неоднорідними наборами даних, що робить його придатним для аналізу комплексної медичної інформації, яка включає як клінічні, так і когнітивні, демографічні та поведінкові характеристики пацієнтів.

XGBoost базується на ансамблевому підході, де нові дерева послідовно додаються до моделі з метою корекції похибок попередніх, що дозволяє зосереджуватись на найскладніших для класифікації випадках. Завдяки цьому

модель ефективно виявляє складні нелінійні взаємозв'язки між ознаками, що є надзвичайно важливим у контексті мультифакторної природи нейродегенеративних захворювань. Алгоритм також підтримує вбудовану регуляризацію, що зменшує ризик перенавчання, та має механізми автоматичного оброблення пропущених значень, що полегшує роботу з неповними медичними даними.

Крім того, XGBoost забезпечує високу обчислювальну ефективність і масштабованість, що дозволяє застосовувати його для аналізу великих вибірок, зокрема в умовах клінічних досліджень або епідеміологічного нагляду. Завдяки підтримці метрик важливості ознак, модель також може бути корисною для виявлення ключових факторів ризику розвитку хвороби Альцгеймера, що сприяє не лише точнішій діагностиці, а й формуванню профілактичних рекомендацій.

Таким чином, XGBoost є надзвичайно перспективним алгоритмом для побудови високоточних прогностичних моделей у галузі медицини, зокрема для вчасного виявлення когнітивних порушень та оцінки ризику виникнення хвороби Альцгеймера.

2.3.4 Support Vector Machine

Визначення Support Vector Machine

Support Vector Machine (SVM) — це алгоритм керованого машинного навчання, який застосовується як для задач класифікації, так і для регресійного аналізу. Проте, найчастіше SVM використовується саме для розв'язання задач бінарної класифікації, де він демонструє високу ефективність і точність.

Основна ідея алгоритму полягає у знаходженні оптимальної гіперплощини в N -вимірному просторі, яка максимально розділяє об'єкти різних класів. Гіперплощина визначається таким чином, щоб максимізувати відстань (маржу) між найближчими до неї точками різних класів — так званими опорними векторами (рис. 2.6).

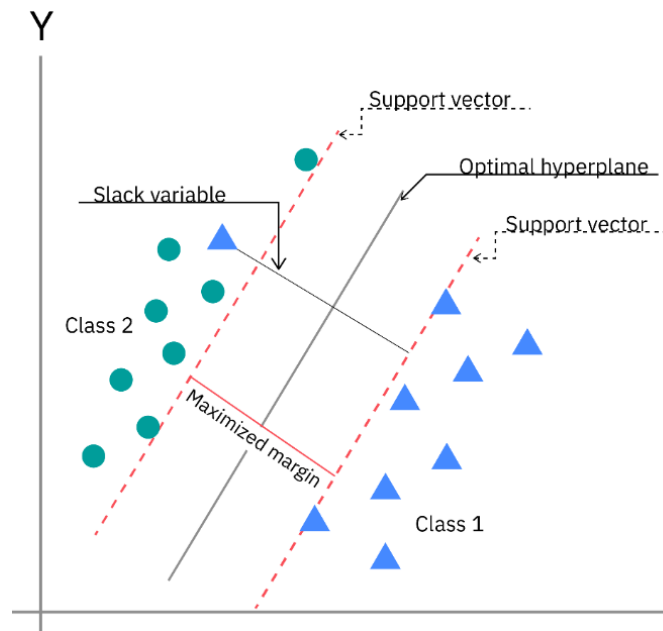


Рисунок 2.6 – Алгоритм Support Vector Machine [29]

Існують два основних типи алгоритмів SVM.

1. Лінійна SVM застосовується у випадках, коли дані є лінійно роздільними, тобто класи можна розділити однією гіперплощиною (у двовимірному випадку — прямою лінією). Така ситуація трапляється рідко у практичних задачах.
2. Нелінійна SVM використовується тоді, коли дані не є лінійно роздільними, що є типовим для реальних медичних та біомедичних даних. У таких випадках застосовується метод ядрових функцій (kernel trick), який дозволяє перетворити вхідний простір ознак у простір більшої розмірності, в якому дані можуть стати лінійно роздільними.

Математичне обґрунтування

Розглянемо задачу бінарної класифікації з двома класами, позначеними як +1 і -1. Припустимо, що маємо навчальний набір, який містить вхідні вектори ознак X та відповідні їм мітки класів Y .

Рівняння для лінійної гіперплощини можна записати так:

$$\omega^T x + b = 0$$

де ω — вектор нормалі до гіперплощини, b — це зміщення, що представляє відстань гіперплощини від початку координат вздовж вектора нормалі ω .

Відстань від точки даних до гіперплощини

Відстань між точкою даних x_i та межею прийняття рішення можна розрахувати за формулою:

$$d_i = \frac{\omega^T x_i + b}{\|\omega\|}$$

де $\|\omega\|$ представляє евклідову норму вагового вектора ω .

Лінійний SVM-класифікатор

Відстань від точки даних до гіперплощини:

$$\hat{y} = \begin{cases} 1, & \omega^T x + b \geq 0 \\ 0, & \omega^T x + b < 0 \end{cases}$$

де $\hat{y}$ — це передбачувана мітка точки даних.

Проблема оптимізації для SVM

Для лінійно роздільного набору даних метою є пошук гіперплощини, яка максимізує межу між двома класами, забезпечуючи при цьому правильну класифікацію всіх точок даних. Це призводить до наступної задачі оптимізації:

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2$$

З урахуванням обмеження:

$$y_i(\omega^T x_i + b) \geq 1, \quad i = 1, 2, 3, \dots, m$$

де y_i — це мітка класу (+1 або -1) для кожного навчального екземпляра, x_i — вектор ознак для i -го навчального екземпляра, m — загальна кількість навчальних екземплярів.

Умова $y_i(\omega^T x_i + b) \geq 1$ гарантує, що кожна точка даних правильно класифікована та знаходиться за межами поля (margin).

Переваги Support Vector Machine

Алгоритм Support Vector Machine (SVM) має низку важливих переваг, що зумовлюють його ефективність у вирішенні складних задач класифікації, зокрема у медичній сфері. Насамперед, SVM ефективно працює з даними складної структури, де межі між класами є нечіткими або нелінійними. Завдяки підтримці різноманітних ядерних функцій (лінійних, поліноміальних, радіально-базисних), модель може адаптуватися до специфіки конкретного набору даних та виявляти складні залежності між ознаками. Алгоритм також демонструє високу продуктивність у високорозмірних просторах, де кількість ознак перевищує кількість прикладів. Однією з ключових властивостей SVM є низька схильність до перенавчання, оскільки метод спрямований на максимізацію маржі між класами, що сприяє формуванню узагальненої моделі. Крім того, алгоритм стійкий до викидів, оскільки орієнтується лише на опорні вектори — приклади, що найбільше впливають на положення гіперплощини розділення.

Недоліки Support Vector Machine

Попри численні переваги, алгоритм Support Vector Machine (SVM) має низку суттєвих обмежень, які необхідно враховувати при його застосуванні. Зокрема, SVM може бути обчислювально затратним та повільним у процесі навчання, особливо при роботі з великими обсягами даних, що обмежує його застосування у задачах із великою кількістю прикладів. Крім того, модель потребує ретельного налаштування гіперпараметрів, зокрема параметрів регуляризації та функції ядра, а також вибору типу ядра. Цей процес часто

вимагає багатьох ітерацій та застосування методів автоматизованої оптимізації, таких як крос-валідація чи пошук сіткою. У випадках, коли дані мають значний рівень шуму або перетинання між класами, SVM може демонструвати нестабільність і втрачати точність, оскільки намагається знайти чітке розділення навіть там, де воно відсутнє. Ще одним недоліком є обмежена інтерпретованість моделі, особливо при використанні нелінійних ядер, що ускладнює пояснення логіки прийняття рішень — важливий аспект у медичних застосуваннях. Окрім цього, модель використовує багато пам'яті, оскільки зберігає всі опорні вектори, що може призводити до значного споживання ресурсів при масштабуванні на великі набори даних. Таким чином, хоча SVM є потужним інструментом, його ефективне використання вимагає зваженого підходу до підготовки даних, вибору гіперпараметрів та обсягу обчислювальних ресурсів.

Придатність для прогнозування ризику виникнення захворювання Альцгеймера

Алгоритм Support Vector Machine (SVM) є доволі ефективним методом керованого машинного навчання для задач класифікації, що робить його перспективним інструментом у медичній діагностиці, зокрема для прогнозування ризику виникнення хвороби Альцгеймера. Основною перевагою SVM є здатність ефективно працювати у високорозмірних просторах ознак, що є характерним для медичних даних, які можуть містити численні клінічні, когнітивні, демографічні та поведінкові змінні.

Алгоритм SVM побудований на принципі максимізації маржі між класами, що забезпечує високу узагальнюючу здатність моделі та стійкість до перенавчання, особливо у випадках з обмеженим обсягом вибірки. Ця властивість є важливою при роботі з медичними даними, де розмір вибірки може бути обмежений, а ознак — багато. Крім того, SVM може моделювати як лінійні, так і складні нелінійні залежності завдяки використанню ядрових функцій (kernel trick). Це дозволяє моделі адаптуватися до гетерогенної

структури даних, виявляючи приховані закономірності між ознаками, що пов'язані з розвитком когнітивних порушень.

У контексті прогнозування ризику хвороби Альцгеймера, SVM може бути використаний для класифікації пацієнтів за наявністю або відсутністю когнітивних симптомів, а також для ідентифікації підгруп із підвищеним ризиком на основі аналізу багатовимірних ознак. Завдяки фокусуванню на опорних векторах, тобто найбільш інформативних зразках даних, модель формує стійкі межі між класами, що є важливим у задачах, де спостерігається часткове перекриття ознак або наявність шуму.

Таким чином, модель Support Vector Machine є доцільною та ефективною для побудови діагностичних систем на основі машинного навчання, орієнтованих на раннє виявлення та прогнозування ризику розвитку хвороби Альцгеймера. Її здатність працювати з високорозмірними, нелінійними та частково нечіткими даними дозволяє отримувати точні та надійні результати, що можуть бути використані як основа для прийняття клінічних рішень.

2.3.5 Stacking Ensemble

Визначення Stacking

Stacking — це сучасна техніка ансамблевого навчання, яка перевершує традиційні методи, такі як bagging та boosting, завдяки використанню мультирівневої архітектури моделей. Основна ідея стекінгу полягає не лише у поєднанні вихідних прогнозів кількох базових моделей, а й у навчанні мета-моделі (рівень-1), яка здійснює фінальне рішення на основі результатів базових моделей (рівень-0).

Архітектура стекінгу, зображена на рис. 2.7, передбачає два рівні навчання. На першому рівні функціонують базові моделі, які навчаються незалежно на одній і тій самій навчальній вибірці. До базових моделей можуть належати будь-які класифікатори — наприклад, Decision Tree, Random Forest,

Support Vector Machine. Кожна з цих моделей має свою специфіку побудови гіпотези та здатна фіксувати різні аспекти закономірностей у даних. На другому рівні застосовується мета-модель, яка не використовує вихідні ознаки, а навчається виключно на виходах базових моделей. Метою мета-моделі є оптимальне агрегування отриманих прогнозів та підвищення загальної точності системи.

У ролі мета-моделі часто використовують лінійну або логістичну регресію, які мають здатність виявляти залежності між виходами моделей першого рівня та істинними мітками. Таким чином, стекінг дозволяє інтегрувати сильні сторони різних алгоритмів, зменшити їхні індивідуальні похибки та підвищити точність прогнозування.

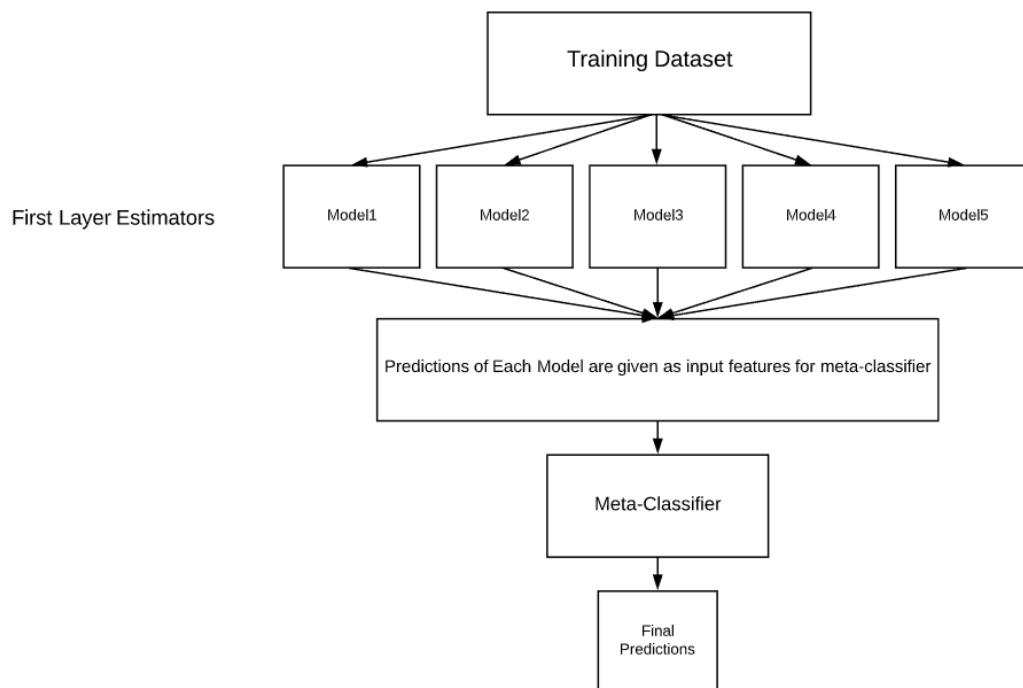


Рисунок 2.7 - Структура Stacking [30]

У контексті прогнозування ризику виникнення хвороби Альцгеймера, стекінг є надзвичайно ефективним, оскільки дозволяє комбінувати моделі з різними властивостями. Такий підхід сприяє створенню більш надійної, стійкої та високоточної системи прогнозування, що особливо важливо для

задач медичної діагностики, де помилка моделі може мати критичні наслідки для пацієнта.

Математичне обґрунтування

Метод Stacking є ансамблевим підходом до машинного навчання, який поєднує декілька базових моделей (першого рівня) за допомогою мета-моделі (другого рівня), що забезпечує формування узагальненого прогнозу на основі проміжних результатів. На відміну від традиційних методів ансамблювання, таких як Bagging чи Boosting, стекінг дозволяє поєднувати моделі різної природи, наприклад, дерева рішень, SVM, нейронні мережі, що підвищує гнучкість і здатність до адаптації до складних структур даних.

Нехай задано вхідний вектор ознак X , а також M базових моделей $f_1(x), f_1(x), \dots, f_M(x)$, кожна з яких формує власний прогноз. У моделі стекінгу ці прогнози передаються як вхідні дані до мета-моделі $g()$, яка генерує остаточне передбачення:

$$\hat{f}(x) = g(f_1(x), f_1(x), \dots, f_M(x))$$

Така формалізація дозволяє мета-моделі навчатися на узагальнених ознаках, які відображають різні підходи до інтерпретації вхідних даних, і, відповідно, зменшувати як дисперсійну, так і зміщену складову похибки (MSE), особливо якщо функція $g()$ є нелінійною. Це дозволяє моделі стекінгу досягати високої узагальнюючої здатності навіть при обробці складних і неоднорідних медичних даних.

Процес навчання моделі стекінгу включає кілька ключових етапів.

1. Навчання базових моделей.

Обрані моделі першого рівня навчаються на навчальній вибірці. З метою забезпечення різноманітності кожна модель може мати різну архітектуру або набір гіперпараметрів.

2. Прогнозування на валідаційній вибірці.

Після навчання кожна з базових моделей використовується для прогнозування міток на окремій валідаційній підмножині.

3. Побудова мета-моделі.

Створюється модель другого рівня, яка навчається на вихідних прогнозах базових моделей. Тобто прогноз кожної базової моделі слугує ознакою для навчання мета-рівня.

4. Формування остаточного прогнозу.

Під час тестування вхідні дані проходять через усі базові моделі, їхні прогнози передаються на вхід мета-моделі, яка формує остаточне рішення.

Таким чином, стекінг дозволяє зменшити загальну похибку передбачення за рахунок поєднання сильних сторін окремих моделей, при цьому забезпечуючи гнучкість у виборі архітектури ансамблю. Це робить метод надзвичайно перспективним для задач медичного прогнозування, зокрема для оцінки ризику виникнення хвороби Альцгеймера, де важливо забезпечити високу точність і узагальнювальну здатність моделей в умовах багатовимірних, змішаних та часто неповних даних.

Переваги Stacking

Метод Stacking має низку важливих переваг, які зумовлюють його ефективність у складних задачах прогнозування. Однією з ключових переваг є підвищення точності моделей, оскільки об'єднання прогнозів кількох базових алгоритмів дозволяє сформувати більш надійний та узагальнений вихідний результат. Стекінг забезпечує синергетичний ефект завдяки поєднанню різнорідних моделей, таких як дерева рішень, логістична регресія чи метод опорних векторів, кожна з яких робить внесок до загального прогнозу, враховуючи свої специфічні сильні сторони. Ще однією суттєвою перевагою є зменшення ризику перенавчання, особливо при використанні крос-валідації на етапі формування ознак для мета-моделі, що сприяє

збалансуванню похибок окремих моделей. Крім того, мета-модель у стекінгу має здатність навчатися на помилках базових моделей, що дозволяє їй адаптивно коригувати кінцевий результат. Важливо й те, що стекінг є високофлексивним підходом, який дозволяє довільно обирати комбінації моделей нульового та першого рівнів залежно від характеру даних і специфіки задачі. У сукупності ці характеристики роблять стекінг потужним і універсальним методом ансамблювання, що особливо цінно в медичних застосуваннях, де точність і надійність прогнозу є критично важливими.

Недоліки Stacking

Попри високу ефективність, метод Stacking має низку обмежень, які необхідно враховувати при його використанні. Насамперед, цей підхід є структурно складнішим у реалізації порівняно з базовими моделями або іншими ансамблевими методами, оскільки передбачає побудову кількох рівнів моделей та потребує ретельного налаштування. Крім того, стекінг є обчислювально затратним, оскільки потребує навчання як кількох базових моделей, так і мета-моделі, що зумовлює триваліший час навчання та вищі ресурси. Ще одним недоліком є низька інтерпретованість моделі, адже складна структура з кількох рівнів ускладнює пояснення кінцевого прогнозу, що може бути критично важливо в медичних застосуваннях. Також існує ризик перенавчання, особливо у випадках використання великої кількості складних мета-моделей. Нарешті, стекінг вимагає достатньо великого обсягу даних для ефективного навчання як базових моделей, так і мета-рівня, що може бути проблематичним у задачах з обмеженою кількістю якісних медичних спостережень.

Придатність для прогнозування ризику виникнення захворювання Альцгеймера

Метод Stacking є одним з найпотужніших технік ансамблевого навчання, який демонструє високу ефективність у задачах класифікації та

прогнозування, зокрема у медичних застосуваннях. Його сутність полягає у поєднанні кількох базових моделей з використанням мета-моделі, яка здійснює остаточне передбачення на основі проміжних прогнозів базових алгоритмів. Такий підхід дозволяє інтегрувати переваги різних моделей машинного навчання в одну комбіновану структуру, здатну покращити загальну точність прогнозування.

У контексті прогнозування ризику виникнення хвороби Альцгеймера, де дані є складними, різнотипними та багатовимірними, стекінг дозволяє поєднувати моделі з різною чутливістю до окремих ознак. Наприклад, дерева рішень добре обробляють категоріальні змінні та виявляють ієрархічні закономірності, SVM ефективні у високорозмірних просторах, а XGBoost добре справляється з неоднорідністю даних. Об'єднання таких моделей у рамках стекінгу дозволяє компенсувати їхні індивідуальні слабкі сторони та сформувати більш стійкий до шуму та недосконалостей даних класифікатор.

Ключовою перевагою стекінгу є здатність знижувати як зміщення, так і дисперсію моделі, що є критичним для отримання надійних медичних прогнозів. Мета-модель, навчаючись на виходах базових моделей, виявляє закономірності у їхніх помилках та узгодженостях, що дозволяє коригувати помилки і підвищувати узагальнюючу здатність. У поєднанні з крос-валідацією під час формування навчального набору для мета-рівня це забезпечує стійкість до перенавчання та підвищує достовірність результатів.

Таким чином, модель Stacking є доцільною та високоефективною для задач прогнозування ризику виникнення захворювання Альцгеймера, оскільки забезпечує гнучке поєднання різних моделей, високу точність класифікації та стійкість до складних і неоднорідних вхідних даних, що характерні для реальних медичних сценаріїв. Її використання сприяє створенню більш надійних інтелектуальних систем підтримки прийняття клінічних рішень.

2.4 Висновки до розділу 2

У цьому розділі здійснюється ґрунтовний аналіз машинного навчання як сучасного інструмента для виявлення закономірностей і прогнозування ризику виникнення захворювання Альцгеймера. Представлено теоретичні засади машинного навчання, математичні підходи, алгоритмічні основи та особливості їхнього застосування в галузі медичної діагностики.

Розділ починається з визначення основних понять і термінів, що використовуються у сфері машинного навчання. Пояснюються типи навчання (кероване, некероване, напівкероване, з підкріпленням) та їх відповідність до задач медичної класифікації. Окрема увага приділяється математичним основам, зокрема ролі лінійної алгебри, математичного аналізу, теорії ймовірностей, математичної статистики, теорії оптимізації, алгоритмам та структурам даних. Це дозволяє розкрити фундаментальні принципи, на яких ґрунтується побудова моделей машинного навчання, а також обґрунтувати їхню здатність до моделювання складних когнітивних процесів і медичних станів.

У межах цього розділу проаналізовано п'ять алгоритмів, обраних для реалізації системи прогнозування: Decision Tree, Random Forest, XGBoost, Support Vector Machine, Stacking Ensemble. Для кожного з них представлено формальні математичні моделі, принципи роботи, переваги та обмеження, а також обґрунтовано їхню доцільність у застосуванні до даних пацієнтів із ризиком виникнення захворювання Альцгеймера. Також розглянуто питання продуктивності, інтерпретованості, складності налаштування моделей та їх адаптивності до медичних даних, що містять демографічні характеристики, когнітивні тести, медичну історію, клінічні показники та інші важливі параметри.

Загалом, цей розділ забезпечує цілісне розуміння математичних і алгоритмічних засад машинного навчання як інструменту діагностики хвороби Альцгеймера та формує основу для практичної реалізації моделей прогнозування ризику даного захворювання.

РОЗДІЛ 3 РЕАЛІЗАЦІЯ ПРОГРАМНОГО ПРОДУКТУ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

У цьому розділі буде описано програмну реалізацію моделей машинного навчання для прогнозування ризику виникнення захворювання Альцгеймера та детально розібрано всі етапи від попередньої обробки даних до навчання моделі та аналізу результатів, вибір інструментів і технологій. Також буде зроблено висновок про доречність використання даних методів машинного навчання для розв’язання поставленої задачі.

3.1 Огляд набору даних

У якості робочого датасету для проведення порівняння ефективності обраних методів у задачі прогнозування ризику виникнення захворювання Альцгеймера було обрано набір даних Alzheimer's Disease Dataset. Цей набір даних містить розширену інформацію про здоров'я 2149 пацієнтів, кожен з яких має унікальні ідентифікатори. Набір даних включає демографічні дані, фактори способу життя, історію хвороби, клінічні вимірювання, когнітивні та функціональні оцінки, симптоми та діагноз хвороби Альцгеймера. Набір даних складається з різноманітних ознак, наведемо ключові з них.

1. Вік (Age)

Ця ознака представляє вік пацієнтів у роках. У вибірці вік коливається від 60 до 90 років. Вік є одним із найсуттєвіших факторів ризику розвитку хвороби Альцгеймера, оскільки ймовірність її виникнення зростає з віком.

2. Стать (Gender)

Ця ознака є бінарною змінною, що вказує на стать пацієнта: значення 0 відповідає чоловічій статі, а 1 — жіночій. Стать може впливати на схильність до певних захворювань, зокрема хвороби Альцгеймера.

3. Рівень освіти (EducationLevel)

Ця категоріальна ознака відображає освітній рівень пацієнтів і ранжується від загальної відсутності освіти до отримання вищої освіти. Рівень освіти асоціюється з когнітивним резервом і є потенційним захисним фактором проти когнітивних порушень.

4. ІМТ (BMI)

Ця ознака є числовим вектором, що відображає індекс маси тіла пацієнтів у діапазоні від 15 до 40. ІМТ використовується як показник ожиріння або недостатньої маси тіла, що може впливати на загальний стан здоров'я.

5. Куріння (Smoking)

Бінарна ознака, що вказує на статус куріння: 0 — не курить, 1 — курить. Куріння пов'язане з окислювальним стресом і судинними порушеннями, що можуть сприяти розвитку нейродегенеративних процесів, зокрема хвороби Альцгеймера.

6. Споживання алкоголю (AlcoholConsumption)

Ця числова ознака відображає кількість алкоголю, споживаного пацієнтом протягом тижня (0–20 одиниць). Надмірне споживання алкоголю може негативно впливати на когнітивні функції та пришвидшувати нейродегенеративні процеси.

7. Фізична активність (PhysicalActivity)

Ця числова ознака вказує на кількість годин фізичної активності на тиждень (від 0 до 10). Регулярна фізична активність асоціюється з кращим станом когнітивного здоров'я та зменшеним ризиком розвитку деменції.

8. Якість дієти (DietQuality)

Ця числова ознака відображає якість харчування за шкалою від 0 до 10. Раціон, багатий на антиоксиданти, омега-3 жирні кислоти та вітаміни, є захисним чинником проти хвороби Альцгеймера.

9. Якість сну (SleepQuality)

Ця ознака відображає оцінку якості сну за шкалою від 4 до 10. Порушення сну, особливо фрагментований сон і апное, можуть бути як симптомом, так і фактором ризику розвитку хвороби Альцгеймера.

10. Сімейний анамнез хвороби Альцгеймера (FamilyHistoryAlzheimers)

Ця бінарна ознака вказує на наявність (1) або відсутність (0) випадків хвороби Альцгеймера серед родичів. Наявність такого анамнезу свідчить про можливу генетичну схильність до захворювання.

11. Серцево-судинні захворювання (CardiovascularDisease)

Ця ознака вказує на наявність серцево-судинних захворювань: 1 — так, 0 — ні. Порушення кровопостачання мозку можуть пришвидшити розвиток судинної деменції або погіршити перебіг хвороби Альцгеймера.

12. Діабет (Diabetes)

Бінарна ознака, що фіксує наявність діабету: 1 — наявний діабет, 0 — діабет відсутній. Діабет 2 типу пов'язаний з підвищеним ризиком розвитку хвороби Альцгеймера через інсулінорезистентність у мозковій тканині.

13. Депресія (Depression)

Ця бінарна ознака вказує на наявність депресивного розладу. Депресія може бути як раннім симптомом хвороби Альцгеймера, так і самостійним фактором ризику когнітивного зниження.

14. Черепно-мозкова травма (HeadInjury)

Ця ознака фіксує наявність в анамнезі черепно-мозкової травми. Такі травми підвищують ризик розвитку хвороби Альцгеймера через пошкодження нейрональних структур і активацію патологічних білків.

15. Гіпертензія (Hypertension)

Ця бінарна ознака вказує на наявність підвищеного артеріального тиску. Хронічна гіпертензія асоціюється з порушенням мозкової перфузії та підвищеним ризиком розвитку деменції.

16. Дезорієнтація (Disorientation)

Ця ознака свідчить про наявність порушення орієнтації в часі або просторі. Дезорієнтація є типовим симптомом прогресуючої деменції.

17. Зміни особистості (PersonalityChanges)

Ця ознака фіксує наявність змін особистісних рис або поведінки. Подібні зміни часто виникають у пацієнтів із хворобою Альцгеймера навіть на ранніх стадіях.

18. Забудькуватість (Forgetfulness)

Бінарна ознака, що відображає наявність скарг на пам'ять. Забудькуватість — один із перших симптомів хвороби Альцгеймера, особливо в аспекті короткочасної пам'яті.

19. MMSE (MMSE)

Ця числова ознака відображає результат тесту Mini-Mental State Examination, у межах від 0 до 30. Нижчі значення свідчать про наявність когнітивного дефіциту і використовуються для оцінки тяжкості деменції.

20. Систолічний артеріальний тиск (SystolicBP)

Ця числова ознака відображає рівень систолічного артеріального тиску пацієнтів у межах від 90 до 180 мм рт. ст. Довготривала гіпертензія асоціюється з дегенеративними змінами в мозковій тканині та підвищеним ризиком розвитку когнітивних порушень, включно з хворобою Альцгеймера.

21. Діагноз (Diagnosis)

Цільова змінна, яка приймає значення 1 у разі наявності захворювання Альцгеймера у пацієнта і 0 у протилежному випадку.

Перші п'ять рядків датасету зображено на рис. 3.1.

	PatientID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity	DietQuality	...	MemoryComplaints	BehavioralProblems	ADL	Confusion	Disorientation	PersonalityChanges	DifficultyCompletingTasks	Forgetfulness	Diagnosis	DoctorInCharge
0	4751	73	0	0	2	22.92749	0	13.297218	6.327112	1.347214	..	0	0	1.725883	0	0	0	1	0	0	1000Confid
1	4752	89	0	0	0	26.627861	0	4.542524	7.619885	0.518767	..	0	0	2.592424	0	0	0	0	1	0	1000Confid
2	4753	73	0	3	1	17.795882	0	19.555085	7.844988	1.826335	..	0	0	7.119548	0	1	0	1	0	0	1000Confid
3	4754	74	1	0	1	33.800817	1	12.209286	8.428001	7.435604	..	0	1	6.481226	0	0	0	0	0	0	1000Confid
4	4755	89	0	0	0	20.716974	0	18.454356	6.310461	0.795498	..	0	0	0.014691	0	0	1	1	0	0	1000Confid

Рисунок 3.1 - Перші п'ять рядків набору даних Credit Card Fraud Detection

Даний датасет підходить для розробки прогностичних моделей, проведення статистичного аналізу та дослідження складної взаємодії факторів, що сприяють розвитку хвороби Альцгеймера.

3.2 Попередня обробка даних

Описова статистика датасету Alzheimer's Disease Dataset до стандартизації зображена на рис. 3.2.

	count	mean	std	min	25%	50%	75%	max
PatientID	2149.0	5825.000000	620.507185	4751.000000	5288.000000	5825.000000	6362.000000	6899.000000
Age	2149.0	74.908795	8.990221	60.000000	67.000000	75.000000	83.000000	90.000000
Gender	2149.0	0.506282	0.500077	0.000000	0.000000	1.000000	1.000000	1.000000
Ethnicity	2149.0	0.697534	0.996128	0.000000	0.000000	0.000000	1.000000	3.000000
EducationLevel	2149.0	1.286645	0.904527	0.000000	1.000000	1.000000	2.000000	3.000000
BMI	2149.0	27.655697	7.217438	15.008851	21.611408	27.823924	33.869778	39.992767
Smoking	2149.0	0.288506	0.453173	0.000000	0.000000	0.000000	1.000000	1.000000
AlcoholConsumption	2149.0	10.039442	5.757910	0.002003	5.139810	9.934412	15.157931	19.989293
PhysicalActivity	2149.0	4.920202	2.857191	0.003616	2.570626	4.766424	7.427899	9.987429
DietQuality	2149.0	4.993138	2.909055	0.009385	2.458455	5.076087	7.558625	9.998346
SleepQuality	2149.0	7.051081	1.763573	4.002629	5.482997	7.115646	8.562521	9.999840
FamilyHistoryAlzheimers	2149.0	0.252210	0.434382	0.000000	0.000000	0.000000	1.000000	1.000000
CardiovascularDisease	2149.0	0.144253	0.351428	0.000000	0.000000	0.000000	0.000000	1.000000
Diabetes	2149.0	0.150768	0.357906	0.000000	0.000000	0.000000	0.000000	1.000000
Depression	2149.0	0.200558	0.400511	0.000000	0.000000	0.000000	0.000000	1.000000
HeadInjury	2149.0	0.092601	0.289940	0.000000	0.000000	0.000000	0.000000	1.000000
Hypertension	2149.0	0.148906	0.356079	0.000000	0.000000	0.000000	0.000000	1.000000
SystolicBP	2149.0	134.264774	25.949352	90.000000	112.000000	134.000000	157.000000	179.000000
DiastolicBP	2149.0	89.847836	17.592496	60.000000	74.000000	91.000000	105.000000	119.000000
CholesterolTotal	2149.0	225.197519	42.542233	150.093316	190.252963	225.086430	262.031657	299.993352
CholesterolLDL	2149.0	124.335944	43.366584	50.230707	87.195798	123.342593	161.733733	199.965665
CholesterolHDL	2149.0	59.463533	23.139174	20.003434	39.095698	59.768237	78.939050	99.980324
CholesterolTriglycerides	2149.0	228.281496	101.986721	50.407194	137.583222	230.301983	314.839046	399.941862
MMSE	2149.0	14.755132	8.613151	0.005312	7.167602	14.441660	22.161028	29.991381
FunctionalAssessment	2149.0	5.080055	2.892743	0.000460	2.566281	5.094439	7.546981	9.996467
MemoryComplaints	2149.0	0.208004	0.405974	0.000000	0.000000	0.000000	0.000000	1.000000
BehavioralProblems	2149.0	0.156817	0.363713	0.000000	0.000000	0.000000	0.000000	1.000000
ADL	2149.0	4.982958	2.949775	0.001288	2.342836	5.038973	7.581490	9.999747
Confusion	2149.0	0.205212	0.403950	0.000000	0.000000	0.000000	0.000000	1.000000
Disorientation	2149.0	0.158213	0.365026	0.000000	0.000000	0.000000	0.000000	1.000000
PersonalityChanges	2149.0	0.150768	0.357906	0.000000	0.000000	0.000000	0.000000	1.000000
DifficultyCompletingTasks	2149.0	0.158678	0.365461	0.000000	0.000000	0.000000	0.000000	1.000000
Forgetfulness	2149.0	0.301536	0.459032	0.000000	0.000000	0.000000	1.000000	1.000000
Diagnosis	2149.0	0.353653	0.478214	0.000000	0.000000	0.000000	1.000000	1.000000

Рисунок 3.2 - Описова статистика для змінних в датасеті Alzheimer's Disease Dataset до стандартизації

Для проведення первинного аналізу ознак досліджуваного набору даних було використано метод `describe()` із бібліотеки `pandas`, який дозволяє отримати базову статистику по кожній змінній. Результати подано на рис. 3.2. Цей опис включає кількість спостережень (`count`), середнє значення (`mean`), стандартне відхилення (`std`), мінімальне (`min`) та максимальне (`max`) значення, 25%, 50%, 75% — відповідно 1-й, 2-й (медіана) та 3-й квартилі.

Як видно з рис. 3.2, більшість числових ознак мають обмежені діапазони значень, що є типовим для медичних та демографічних даних. Зокрема, вік пацієнтів варіюється від 60 до 90 років, а індекс маси тіла (BMI) – від 15 до 39.99. Ознаки фізичної активності та дієти мають середні значення в межах 0–10, що вказує на помірний рівень здорової поведінки серед пацієнтів. Систолічний артеріальний тиск варіюється від 90 до 179 мм рт. ст., а діастолічний – від 60 до 119 мм рт. ст., що охоплює як нормальні, так і критичні показники. Ознака MMSE демонструє середнє значення 14.75 при можливому максимумі 29.99, що вказує на наявність когнітивних порушень у значної частини вибірки. Бінарні змінні (наприклад, діабет, депресія, гіпертензія, плутанина, дезорієнтація, забудькуватість) мають середні значення в межах 0.1–0.5.

Статистика пропущених значень та кардинальності ознак датасету Alzheimer's Disease Dataset зображена на рис. 3.3.

	Count	Null	Null %	Cardinality
PatientID	2149	0	0.0	2149
Age	2149	0	0.0	31
Gender	2149	0	0.0	2
Ethnicity	2149	0	0.0	4
EducationLevel	2149	0	0.0	4
BMI	2149	0	0.0	2149
Smoking	2149	0	0.0	2
AlcoholConsumption	2149	0	0.0	2149
PhysicalActivity	2149	0	0.0	2149
DietQuality	2149	0	0.0	2149
SleepQuality	2149	0	0.0	2149
FamilyHistoryAlzheimers	2149	0	0.0	2
CardiovascularDisease	2149	0	0.0	2
Diabetes	2149	0	0.0	2
Depression	2149	0	0.0	2
HeadInjury	2149	0	0.0	2
Hypertension	2149	0	0.0	2
SystolicBP	2149	0	0.0	90
DiastolicBP	2149	0	0.0	60
CholesterolTotal	2149	0	0.0	2149
CholesterolLDL	2149	0	0.0	2149
CholesterolHDL	2149	0	0.0	2149
CholesterolTriglycerides	2149	0	0.0	2149
MMSE	2149	0	0.0	2149
FunctionalAssessment	2149	0	0.0	2149
MemoryComplaints	2149	0	0.0	2
BehavioralProblems	2149	0	0.0	2
ADL	2149	0	0.0	2149
Confusion	2149	0	0.0	2
Disorientation	2149	0	0.0	2
PersonalityChanges	2149	0	0.0	2
DifficultyCompletingTasks	2149	0	0.0	2
Forgetfulness	2149	0	0.0	2
Diagnosis	2149	0	0.0	2
DoctorInCharge	2149	0	0.0	1

Рисунок 3.3 - Статистика пропущених значень та кардинальності ознак датасету Alzheimer's Disease Dataset

У цьому наборі даних відсутні пропущені значення, а кількість спостережень для кожної змінної становить 2149. Також наведено кількість

унікальних значень у кожному стовпчику, що визначається за допомогою методу `nunique()` із бібліотеки `pandas`.

Оскільки датасет складається з даних різної природи, вони мають також доволі різні діапазони значень, як зображено на рис. 3.4:

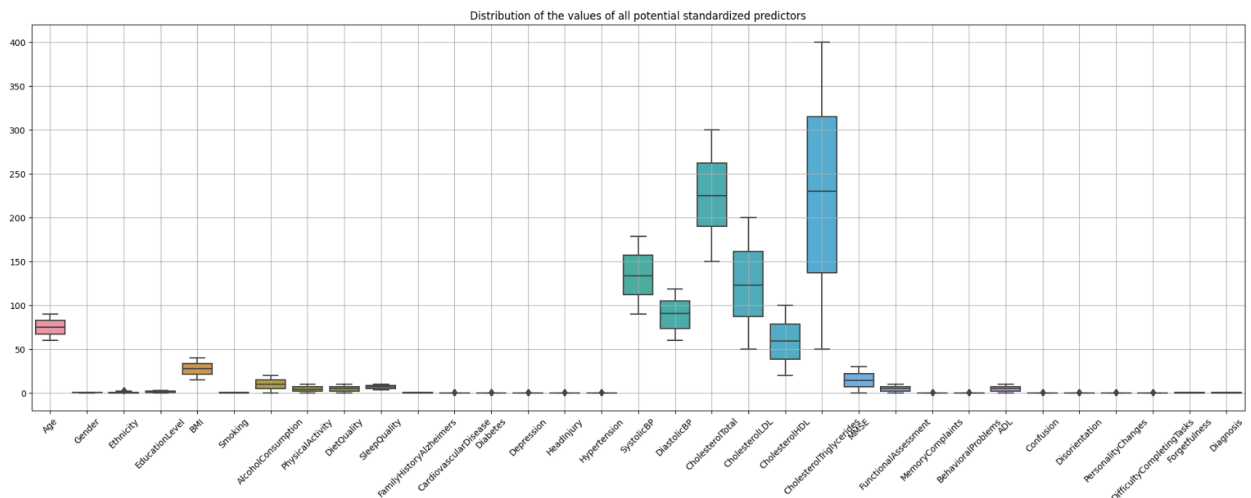


Рисунок 3.4 - Розподіл значень усіх потенційних стандартизованих предикторів

З метою забезпечення коректного функціонування моделей машинного навчання, числові ознаки слід нормалізувати за допомогою стійкого методу масштабування `RobustScaler`, який зменшує вплив викидів, здійснюючи перетворення даних таким чином, щоб центр розподілу відповідав медіані ознаки, а масштаб — міжквартильному розмаху (IQR), що забезпечує стійкість до викидів та нерівномірного розподілу даних.

Для датасету `Alzheimer's Disease Dataset` буде оптимально виконати стандартизацію за допомогою `RobustScaler` з бібліотеки `scikit-learn`. Його використовують, коли дані містять багато викидів, і потрібно підтримувати відносні відстані між точками даних, що не є викидами, або ми працюємо з алгоритмами, чутливими до екстремальних значень. Після застосування `RobustScaler` отримуємо більш зручне подання даних, як показано на рис. 3.5:

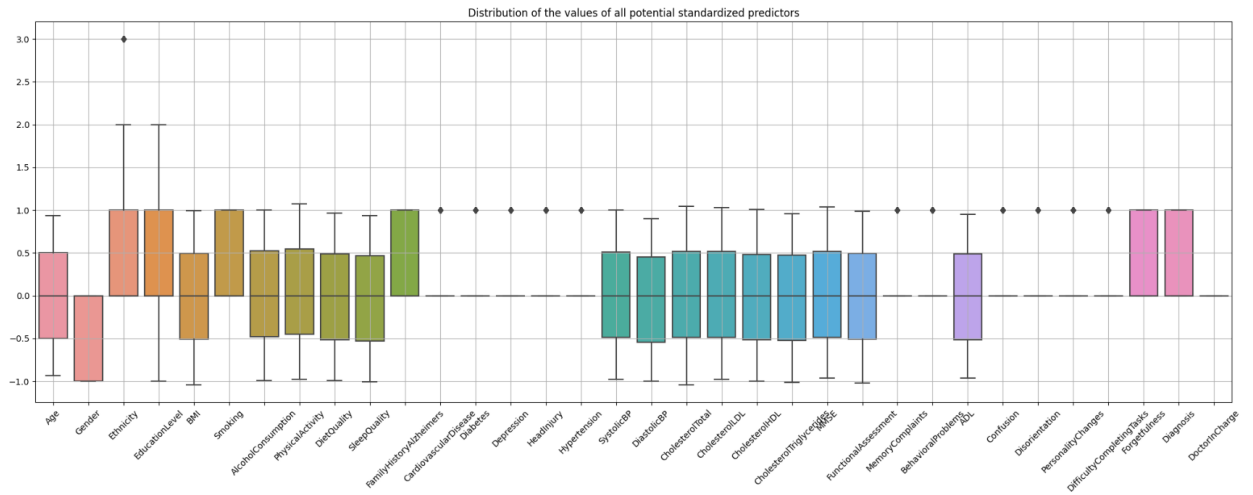


Рисунок 3.5 - Розподіл значень усіх стандартизованих предикторів

Кореляційна структура ознак представлена на рис. 3.6.

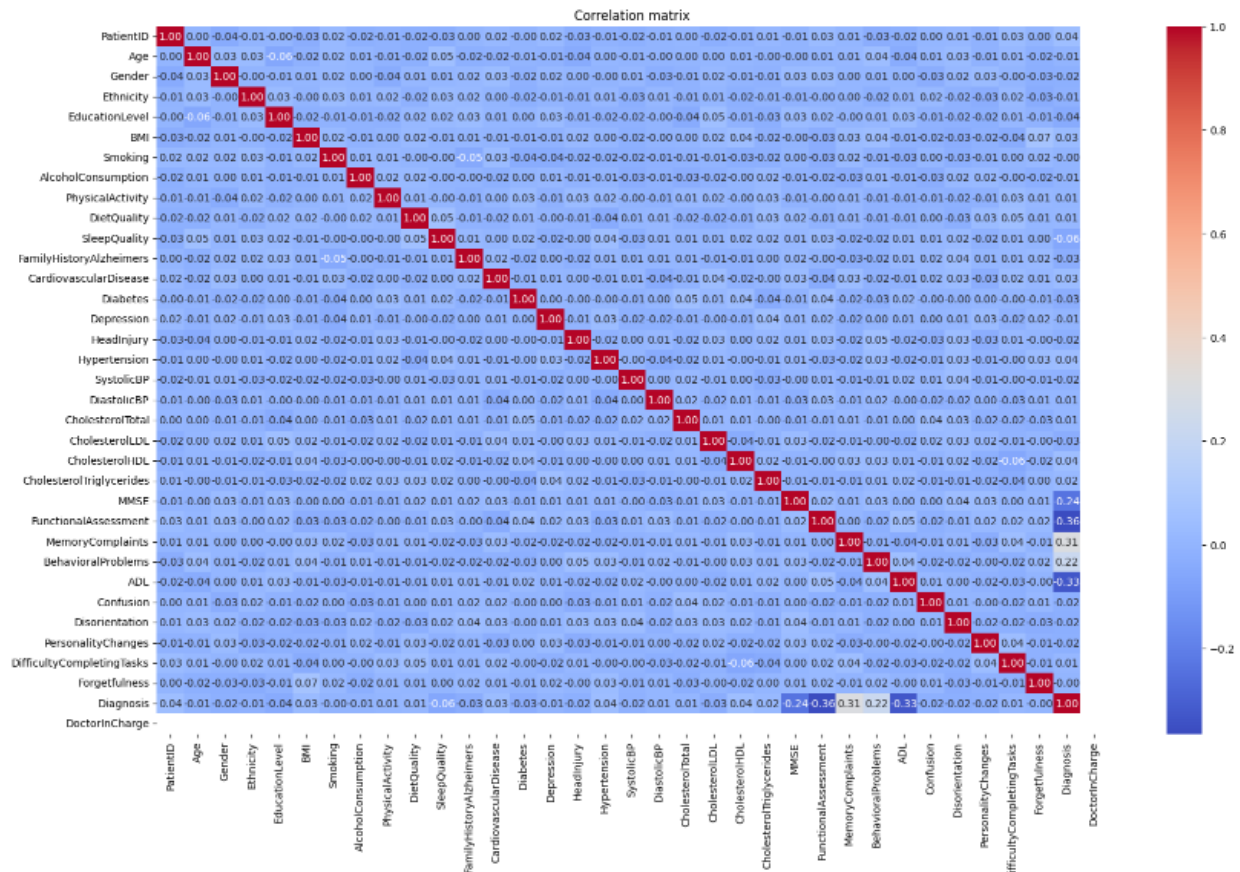


Рисунок 3.6 - Кореляційна матриця набору даних Alzheimer's Disease Dataset

Тобто, ознаки датасету не мають високу кореляцію між собою, а, отже, дані не є мультиколінеарними.

Розподіл ознак Age, EducationLevel, BMI, Smoking, Gender, Diagnosis, MemoryComplaints, Ethnicity, FamilyHistoryAlzheimers зображено на рис. 3.7.

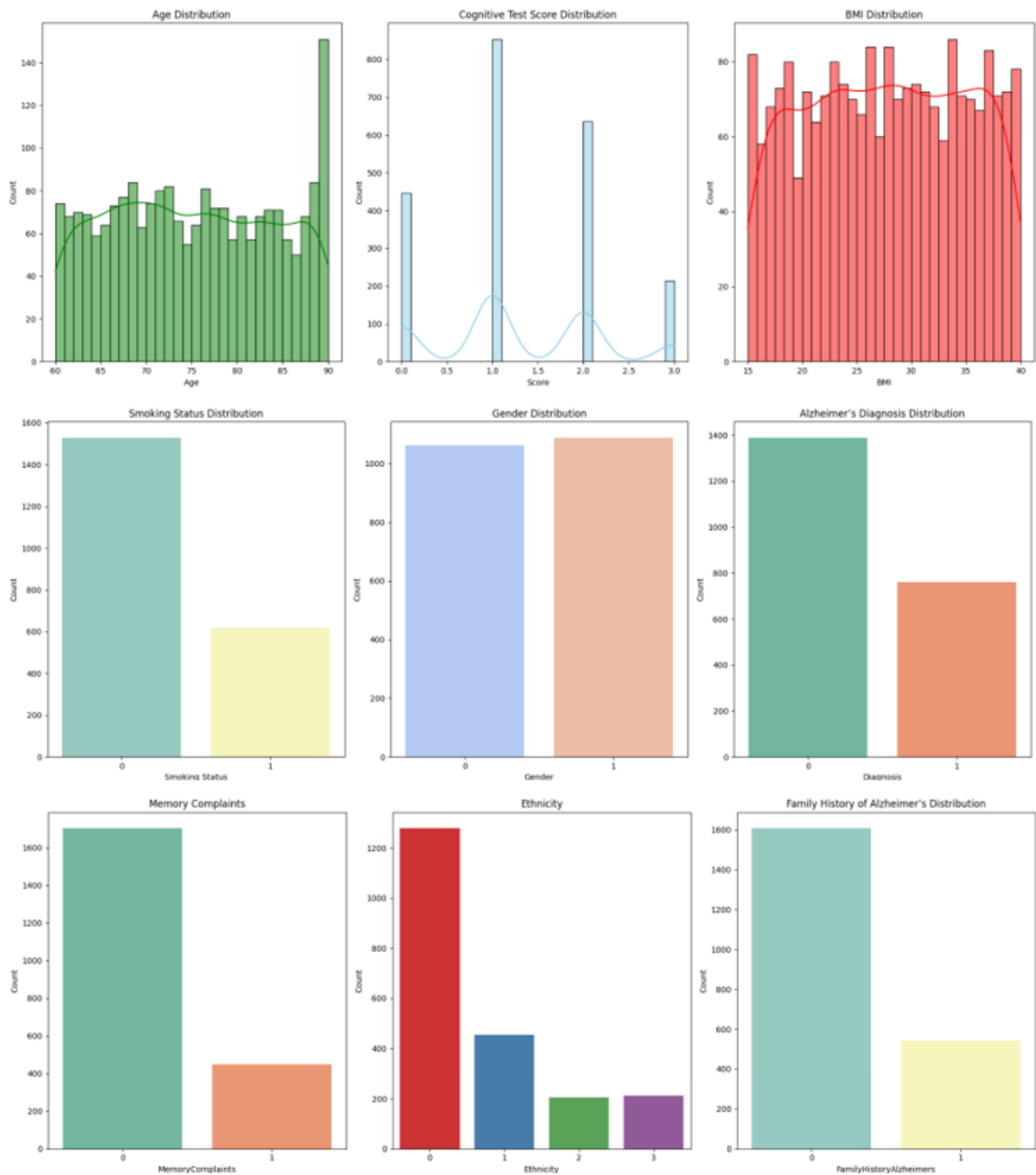


Рисунок 3.7 - Розподіл ознак Age, EducationLevel, BMI, Smoking, Gender, Diagnosis, MemoryComplaints, Ethnicity, FamilyHistoryAlzheimers з датасету Alzheimer's Disease Dataset

3.3 Метрики оцінки отриманих результатів.

Спочатку опишемо метрики, які будуть використані для оцінки якості навчених моделей. Оскільки наша цільова змінна є булевою, існує всього $2 \cdot 2 = 4$ можливих варіанти (табл. 3.1):

1. True Positive - кількість істинно позитивних значень (реально позитивне значення і передбачене - позитивне);
2. True Negative - кількість істинно негативних значень (реально негативне значення і передбачене - негативне);
3. False Positive - кількість невірно позитивних значень (реально негативне значення, але передбачене як позитивне);
4. False Negative - кількість невірно негативних значень (реально позитивне значення, але передбачене як негативне).

Таблиця 3.1 – Можливі категорії бінарної класифікації

Predicted Value	Actual Value		
		positives	negatives
	positives	TP (True Positive)	FP (False Positive)
	negatives	FN (False Negative)	TN (True Negative)

Також визначимо такі метрики, як Accuracy, Precision, Recall і F1-Score.

1. Точність класифікації (Accuracy)

Метрика Accuracy є однією з найпоширеніших та базових метрик для оцінювання якості класифікаційних моделей у машинному навчанні. Точність класифікації визначається як відношення кількості правильних передбачень до загальної кількості прикладів у тестовій вибірці. Для бінарної класифікації

правильними вважаються як позитивні передбачення, коли модель вірно класифікує позитивний клас (True Positives - TP), так і негативні передбачення, коли модель коректно розпізнає негативний клас (True Negatives - TN). Неправильні передбачення поділяються на False Positives (FP) - випадки, коли негативний приклад класифікується як позитивний, та False Negatives (FN) — коли позитивний приклад класифікується як негативний. Формула має наступний вигляд:

$$\text{Accuracy} = \frac{\text{Кількість правильних передбачень}}{\text{Загальна кількість передбачень}} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Точність позитивного класу (Precision)

Метрика Precision є важливою характеристикою ефективності класифікаційних моделей у задачах бінарної або багатокласової класифікації, особливо коли критично важливим є мінімізація кількості хибнопозитивних передбачень. Вона відображає частку дійсно позитивних об'єктів серед усіх передбачених моделлю як позитивні. Дана метрика є особливо релевантною в тих задачах, де вартість хибнопозитивної помилки є вищою, ніж хибнонегативної. Точність позитивного класу визначається як відношення кількості істинно позитивних передбачень (TP) до загальної кількості всіх позитивних передбачень, тобто суми істинно позитивних (TP) і хибнопозитивних (FP) передбачень. Формула має наступний вигляд:

$$\text{Precision} = \frac{\text{Кількість істинно позитивних передбачень}}{\text{Кількість усіх передбачень позитивного класу}} = \frac{TP}{TP + FP}$$

3. Повнота (Recall)

Метрика Recall є однією з ключових метрик для оцінювання якості класифікаційної моделі, особливо в задачах, де критично важливим є виявлення всіх об'єктів певного класу. Вона вимірює здатність моделі

правильно виявляти всі наявні позитивні приклади у вибірці, тобто визначає частку істинно позитивних результатів серед усіх фактичних позитивних об'єктів. Ця метрика має особливу важливість у задачах, де пропущений позитивний випадок (хибнонегативне передбачення) є небажаним або потенційно шкідливим. Повнота визначається як відношення кількості істинно позитивних передбачень (TP) до загальної кількості об'єктів, що насправді належать до позитивного класу, тобто суми істинно позитивних (TP) і хибнонегативних (FN) прикладів. Формула має наступний вигляд:

$$\text{Recall} = \frac{\text{Кількість істинно позитивних передбачень}}{\text{Кількість усіх фактичних позитивних прикладів}} = \frac{TP}{TP + FN}$$

4. F1-міра (F1-Score)

Метрика F1-score є узагальненою характеристикою якості класифікаційної моделі, яка поєднує дві основні метрики — Precision (точність) та Recall (повноту) — у єдине гармонійне середнє значення. Вона особливо корисна у випадках, коли існує необхідність досягти балансу між точністю та повнотою, зокрема у задачах із нерівномірним розподілом класів або за умови високої вартості як хибнопозитивних, так і хибнонегативних передбачень.

Гармонійне середнє, на відміну від арифметичного, зменшується у разі значного дисбалансу між цими показниками, що стимулює досягнення одночасно високих значень як точності, так і повноти.

Формально, F1-score визначається наступним чином:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Підставивши формули для Precision і Recall через показники TP (True Positives), FP (False Positives) та FN (False Negatives), можна представити F1-score також у вигляді:

$$F1\text{-score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

5. Площа під кривою (AUC)

Показник AUC (Area Under the Curve – площа під кривою) – це міра, яка дозволяє підсумовувати продуктивність моделі одним числом, вимірюючи площу під кривою ROC. Крива ROC – це графік, що ілюструє продуктивність класифікаційної моделі за всіх можливих порогів класифікації. Вісь X даного графіка є False Positive Rate (FPR), тобто хибнопозитивну частоту, а вісь Y – True Positive Rate (TPR), тобто істиннопозитивну частоту. Крива ROC є графічним представленням компромісу між чутливістю і специфічністю при різних порогах класифікації. Ідеальна модель класифікації буде прагнути точки у верхньому лівому куті графіка, де TPR дорівнює 1, а FPR дорівнює 0. AUC коливається від 0 до 1, де вище значення AUC вказує на більш високу продуктивність моделі. AUC рівний 0.5 вказує на відсутність дискримінаційної здатності моделі, тоді як AUC рівний 1.0 означає ідеальну відмінність класів. Важливі властивості AUC – це інваріантність до порога класифікації та масштабу передбачень. Інваріантність до масштабу передбачень означає, що AUC не залежить від масштабу ймовірностей, які генерує модель.

3.4 Навчання моделей і аналіз отриманих результатів

У цьому розділі буде описано процес навчання обраних моделей машинного навчання для задачі прогнозування ризику виникнення захворювання Альцгеймера. В ході роботи було здійснено підбір гіперпараметрів із використанням методу GridSearch. Цей метод полягає у переборі всіх можливих комбінацій заданих параметрів у певному діапазоні з

метою знаходження оптимального набору, що забезпечує найкращі показники якості моделі за обраною метрикою. Таким чином, для кожної моделі було знайдено гіперпараметри, що дозволили досягти максимальної ефективності при збереженні здатності до узагальнення. У подальших підрозділах наведено результати оцінювання моделей за допомогою низки метрик класифікації, а також здійснено порівняльний аналіз їхньої продуктивності та доцільності застосування в межах поставленого завдання.

3.4.1 Decision Tree

Спершу наведемо і проаналізуємо таблицю з результуючими метриками моделі наведених в табл. 3.2.

Таблиця 3.2 – Метрики моделі Decision Tree

Accuracy	Precision	Recall	F1-Score	AUC
0.939535	0.940049	0.939535	0.939705	0.936976

Загальна точність класифікації свідчить про те, що модель правильно класифікувала близько 93.95% усіх прикладів. Це є досить високим показником, який демонструє, що модель в цілому виконує своє завдання ефективно. Precision вказує, що серед усіх випадків, які модель класифікувала як позитивні (тобто ризик виникнення хвороби Альцгеймера), приблизно 94% дійсно мали цю ознаку. Це особливо важливо в медичних задачах, оскільки висока точність дозволяє мінімізувати кількість хибнопозитивних результатів (False Positives), що може зменшити кількість непотрібних обстежень або втручань. Recall показує, що з усіх реальних випадків ризику захворювання модель змогла правильно ідентифікувати 93.95%. Високе значення recall є критичним у контексті скринінгу хвороби Альцгеймера, оскільки дозволяє виявити максимальну кількість пацієнтів, що потенційно потребують подальшого медичного обстеження. Значення F1-score (0.939705) вказує на

добре збалансовану роботу моделі, без явного перекосу в бік мінімізації хибнопозитивних або хибнонегативних результатів. Значення AUC рівний 0.936976 свідчить про те, що модель здатна коректно відрізнити випадки високого ризику від випадків низького ризику з імовірністю понад 93%.

Тепер проаналізуємо матрицю невідповідностей (confusion matrix) зображену на рис. 3.8.

Модель допустила лише 15 хибнопозитивних передбачень, що становить дуже малу частку від загальної кількості прикладів, і є позитивним аспектом при скринінгу. Також отримано 11 хибнонегативних передбачень, тобто лише 11 осіб із наявним ризиком не були ідентифіковані. Зважаючи на загальну кількість реально позитивних випадків ($TP + FN = 153$), це становить менше ніж 7.2% помилки, що є прийнятним показником для попередньої діагностики.

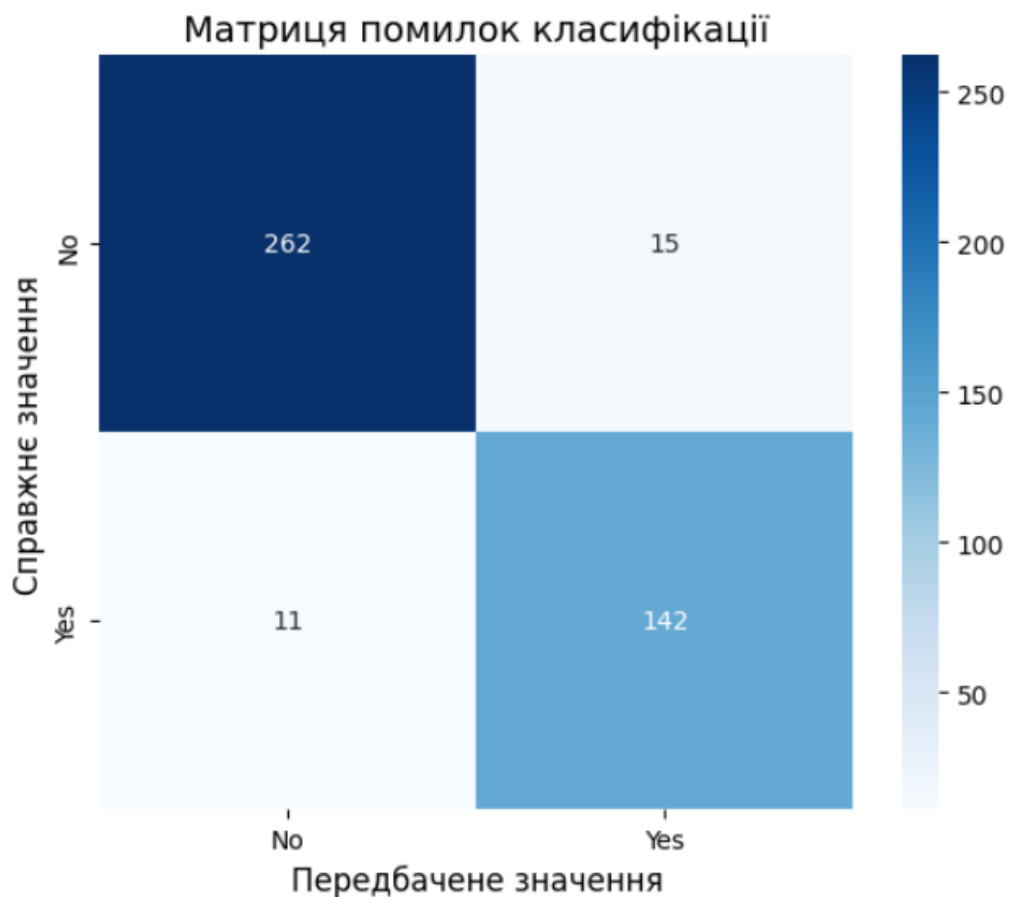


Рисунок 3.8 - Матриця помилок Decision Tree

Отримані результати свідчать про високу якість побудованої моделі дерева рішень у задачі прогнозування ризику виникнення хвороби Альцгеймера. Усі ключові метрики перевищують 0.93, що вказує на стабільну та збалансовану продуктивність моделі як з точки зору виявлення пацієнтів з підвищеним ризиком, так і з точки зору уникнення хибних діагнозів.

3.4.2 Random Forest

Спершу наведемо і проаналізуємо таблицю з результуючими метриками моделі наведених в табл. 3.3.

Таблиця 3.3 – Метрики моделі Random Forest

Accuracy	Precision	Recall	F1-Score	AUC
0.934884	0.936597	0.934884	0.933873	0.915811

Загальна точність моделі свідчить про те, що 93.48% усіх передбачень були правильними. Високе значення Precision означає, що серед усіх випадків, які модель класифікувала як позитивні (наявність ризику захворювання), близько 93.65% виявилися істинно позитивними. Це важливо для уникнення надмірної кількості помилкових медичних втручань у пацієнтів, що насправді не мають ризику. Оскільки Recall дорівнює 0.9348, то модель виявляє приблизно 93.48% усіх справжніх випадків ризику захворювання, що свідчить про її високу чутливість. Такий показник є критичним у контексті медичних застосувань, де пропущені випадки (False Negatives) можуть призвести до відсутності своєчасної діагностики. Значення F1-score вище 0.93 вказує на відмінну узгодженість між precision та recall. Значення AUC рівне 0.9158 вказує на високу дискримінативну здатність і залишається достатньо високим для практичного використання.

Тепер проаналізуємо матрицю невідповідностей (confusion matrix) зображену на рис. 3.9.

Було отримано 5 False Positive передбачень – кількість хибнопозитивних передбачень є дуже низькою, що позитивно впливає на клінічну доцільність моделі, оскільки кількість осіб, яких помилково ідентифікували як хворих, мінімальна. Виявлено дещо більшу кількість помилково пропущених позитивних випадків (False Negative = 23). Це може викликати занепокоєння у контексті медичної діагностики, оскільки такі помилки можуть призводити до недооцінки реального ризику у пацієнтів. Загалом, зі 153 реальних позитивних випадків модель правильно ідентифікувала 130, що становить 85% виявлення.

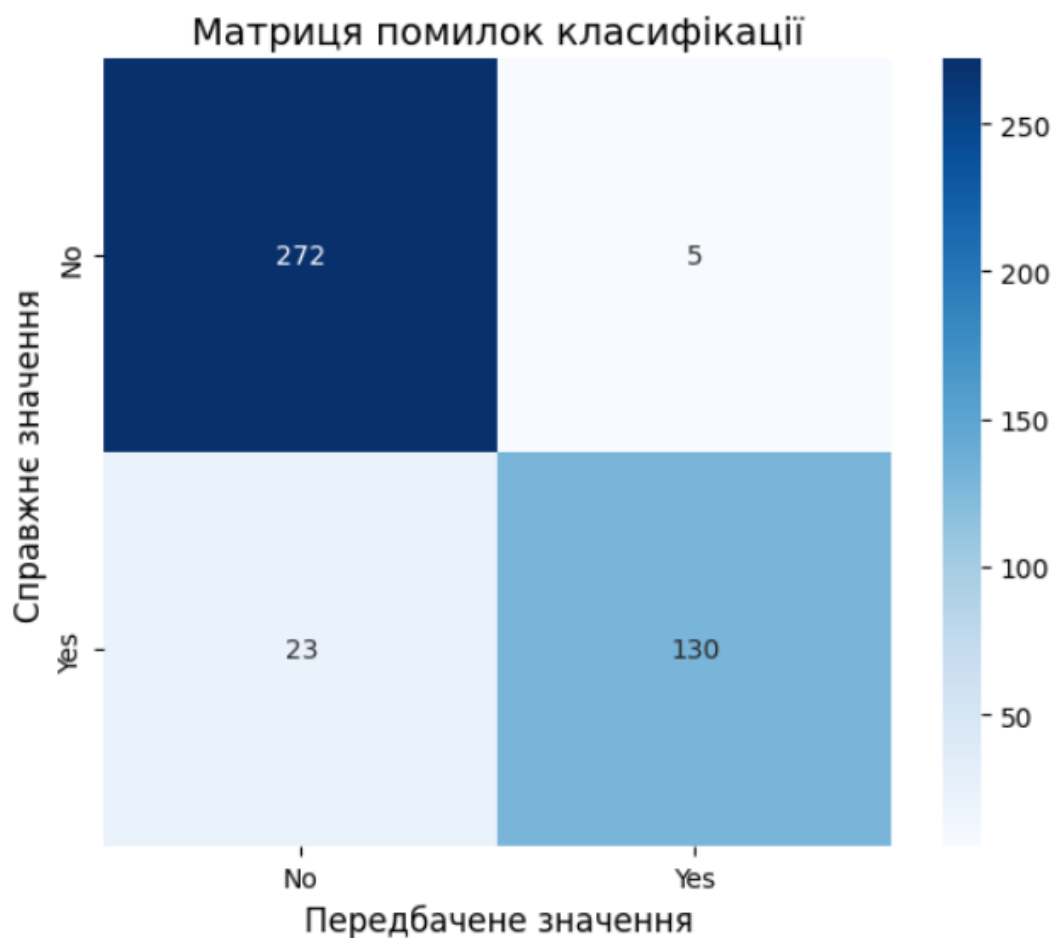


Рисунок 3.9 - Матриця помилок Random Forest

Модель Random Forest демонструє високий рівень точності, надійності та стабільності у завданні класифікації ризику хвороби Альцгеймера. Таким чином, модель Random Forest є доречним вибором для використання як частини інструменту попередньої діагностики або системи підтримки

прийняття рішень у медичній практиці. Вона поєднує високу якість класифікації з кращою генералізацією, що робить її надійною для застосування на різнорідних даних у медичній сфері.

3.4.3 XGBoost

Спершу наведемо і проаналізуємо таблицю з результуючими метриками моделі наведених в табл. 3.4.

Таблиця 3.4 – Метрики моделі XGBoost

Accuracy	Precision	Recall	F1-Score	AUC
0.958140	0.958140	0.958140	0.958140	0.954343

Модель демонструє дуже високу загальну точність, правильно класифікуючи понад 95% прикладів. Precision показує, що понад 95% передбачених випадків ризику дійсно відповідають істинним позитивним прикладам. Такий результат демонструє низький рівень хибнопозитивних передбачень і дозволяє уникнути зайвих діагностичних втручань. Recall дорівнює 0.9581, що свідчить про здатність виявляти переважну більшість пацієнтів, які дійсно мають ризик розвитку хвороби. F1-score рівний 0.9581 підтверджує, що модель добре збалансована і не схильна до перекосів в один із класів. Значення, наближене до одиниці, є ознакою виняткової стабільності моделі. Показник AUC дорівнює 0.9543, що підтверджує, що модель має високий рівень дискримінативної здатності.

Тепер проаналізуємо матрицю невідповідностей (confusion matrix) зображену на рис. 3.10.

Маємо 9 False Positives передбачень — кількість осіб, які були помилково ідентифіковані як такі, що мають ризик хвороби, становить лише 3.2% від усіх негативних випадків. Це надзвичайно низький показник, що знижує ймовірність необґрунтованих клінічних втручань. False Negatives

нараховує 9 передбачень — кількість пацієнтів, у яких модель не виявила наявного ризику, також становить 3.2% від усіх позитивних прикладів. У клінічному контексті це свідчить про високу здатність моделі не пропускати пацієнтів, яким дійсно потрібне подальше медичне обстеження. Загалом, із 153 позитивних прикладів (TP + FN), модель правильно класифікувала 144 (94.1%), що є дуже ефективним результатом.

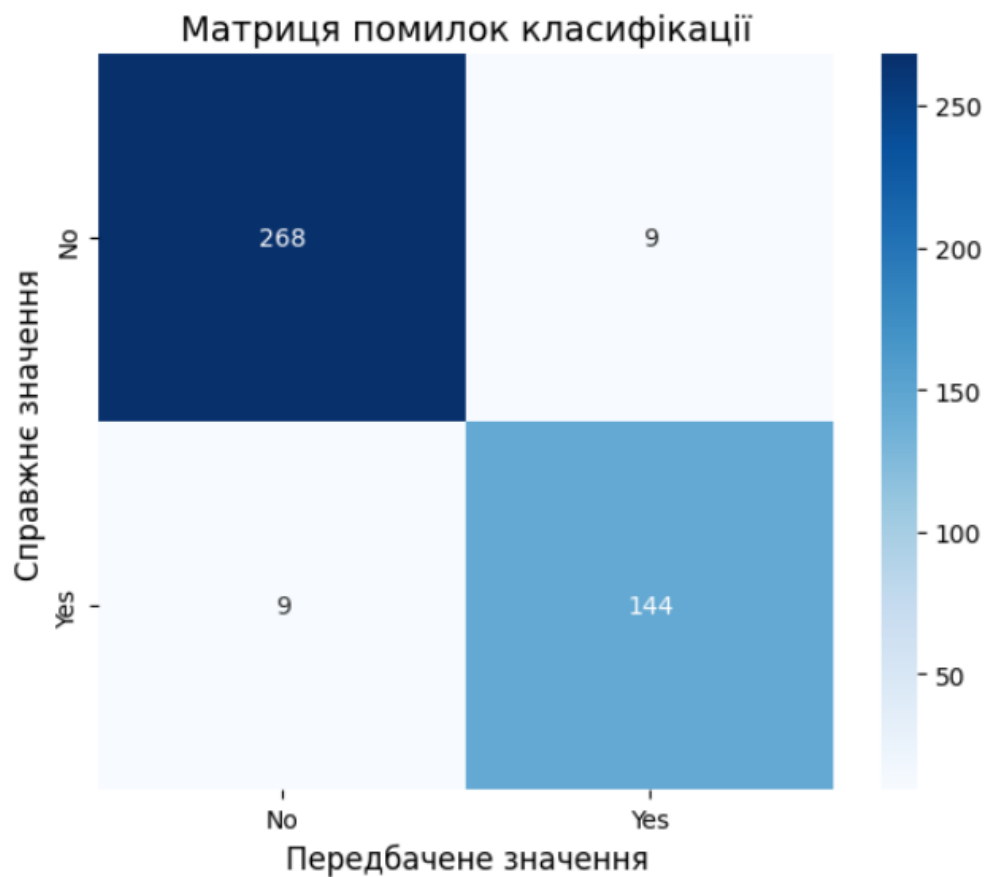


Рисунок 3.10 - Матриця помилок XGBoost

Отже, модель XGBoost є високопродуктивним інструментом для прогнозування ризику виникнення хвороби Альцгеймера, який можна рекомендувати до використання в рамках медичних інформаційних систем, систем раннього виявлення та клінічної підтримки прийняття рішень. Її застосування дозволяє підвищити точність діагностики, мінімізувати кількість хибних діагнозів і своєчасно виявляти осіб із підвищеним ризиком, що відкриває можливість для раннього медичного втручання.

3.4.4 Support Vector Machine

Спершу наведемо і проаналізуємо таблицю з результуючими метриками моделі наведених в табл. 3.5.

Таблиця 3.5 – Метрики моделі Support Vector Machine

Accuracy	Precision	Recall	F1-Score	AUC
0.832558	0.830572	0.832558	0.829960	0.804205

Загальна точність моделі становить приблизно 83.25%, що свідчить про низьку здатність моделі до коректної класифікації об'єктів обох класів. Precision на рівні 83% означає, що з усіх випадків, передбачених як позитивні, близько 17% виявилися хибними. Це вказує на відносно високу кількість False Positive, що в медичному контексті може спричиняти зайві обстеження та психологічне навантаження на здорових пацієнтів. Значення Recall дорівнює 0.8325 також перебуває на середньому рівні, що вказує на те, що понад 16% справжніх позитивних випадків не були виявлені. В умовах ранньої діагностики Альцгеймера така кількість False Negative є критичною, оскільки ці пацієнти залишаються поза увагою клінічного втручання. F1-score рівний 0.8299 свідчить про посередню узгодженість моделі у класифікації позитивного класу. Значення AUC менше 0.81 свідчить про обмежену дискримінативну здатність SVC у цьому випадку.

Тепер проаналізуємо матрицю невідповідностей (confusion matrix) зображену на рис. 3.11.

Маємо 27 False Positives виявлень захворювання у здорових пацієнтів, що становить приблизно 9.8% серед негативних прикладів. Така частота хибнопозитивних передбачень може призводити до непотрібної тривоги серед пацієнтів і до надлишкового навантаження на медичну систему. Також модель не змогла виявити 29.4% (FN = 45) пацієнтів, які справді мали ризик розвитку хвороби. Це найбільш критичний недолік, оскільки пропущені випадки

можуть означати втрату можливості для раннього втручання, що вкрай небажано у сфері охорони здоров'я.

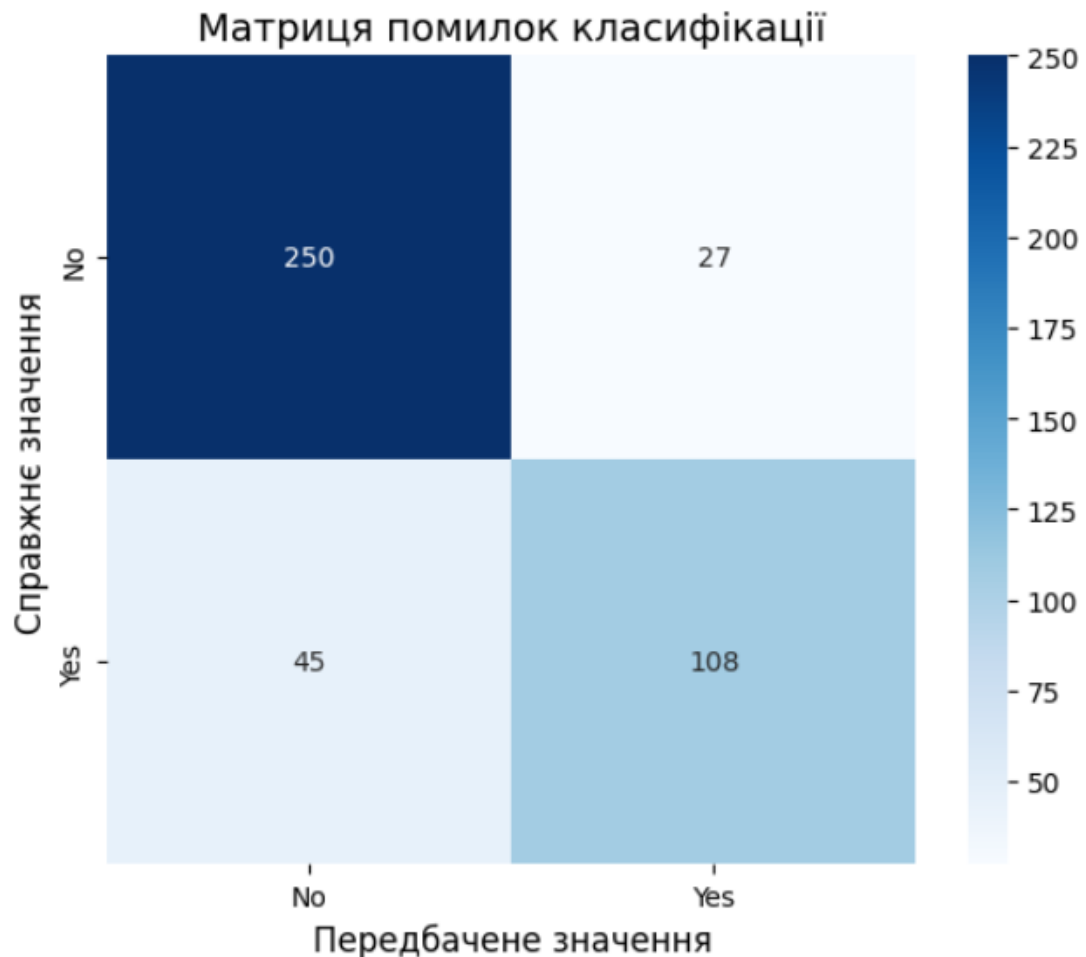


Рисунок 3.11 - Матриця помилок Support Vector Machine

На основі наведених метрик можна дійти висновку, що модель SVC демонструє обмежену ефективність для прогнозування ризику виникнення хвороби Альцгеймера. Її точність, чутливість, і особливо AUC є істотно нижчими у порівнянні з моделями XGBClassifier, Random Forest та Decision Tree. Значна кількість False Negatives (45 випадки) свідчить про те, що ця модель може не виявити велику частку пацієнтів із реальним ризиком, що є неприйнятним у клінічних умовах, де йдеться про своєчасну діагностику та інтервенцію. Враховуючи результати, можна зробити висновок, що модель SVC не є придатною для використання як основний інструмент

діагностичного прогнозування у контексті даної задачі. Її застосування може бути доцільним лише в якості допоміжного методу, або після додаткової оптимізації гіперпараметрів і/або модифікації простору ознак.

3.4.5 Stacking Ensemble

Спершу наведемо і проаналізуємо таблицю з результуючими метриками моделі наведених в табл. 3.6.

Таблиця 3.6 – Метрики моделі Stacking Ensemble

Accuracy	Precision	Recall	F1-Score	AUC
0.962791	0.962791	0.962791	0.962791	0.959416

Ассурасу дорівнює 0.9627 — модель правильно класифікувала майже 96% усіх прикладів у тестовому наборі. Precision показує, що 96.27% передбачених позитивних випадків справді належать до класу ризику. Такий показник свідчить про низький рівень хибнопозитивних передбачень, що є критично важливим у медичній діагностиці, де кожне хибне спрацьовування може спричинити психологічний стрес пацієнта і зайві витрати на діагностичні процедури. Показник Recall рівний 0.9627 перебуває на дуже високому рівні, що вказує на здатність виявляти переважну більшість справжніх випадків ризику. У контексті прогнозування хвороби Альцгеймера така здатність до виявлення навіть слабко виражених випадків є вирішальною для своєчасного втручання. Метрика F1-score дорівнює 0.9627 і вказує на гармонійний баланс між точністю і чутливістю, що є ознакою стабільної та надійної моделі. Площа під ROC-кривою підтверджує високу здатність моделі розрізняти хворих і здорових пацієнтів. AUC, близьке до 0.96, є ще одним свідченням високої дискримінативної здатності моделі, що особливо важливо в медичному контексті.

Тепер проаналізуємо матрицю невідповідностей (confusion matrix) зображену на рис. 3.12.

Було отримано всього 8 False Positives випадків, тобто лише 2.9% серед негативних прикладів були класифіковані помилково як позитивні. Це свідчить про високу специфічність моделі. Також маємо 8 False Negatives передбачень, що становить 2.9% серед позитивних прикладів, які не були виявлені. Це свідчить про надзвичайно високу чутливість моделі, що дозволяє мінімізувати ризик пропущених випадків хвороби. У результаті, із 153 пацієнтів з ризиком розвитку захворювання модель правильно класифікувала 145, що є винятковим результатом (94.8% від позитивних прикладів).

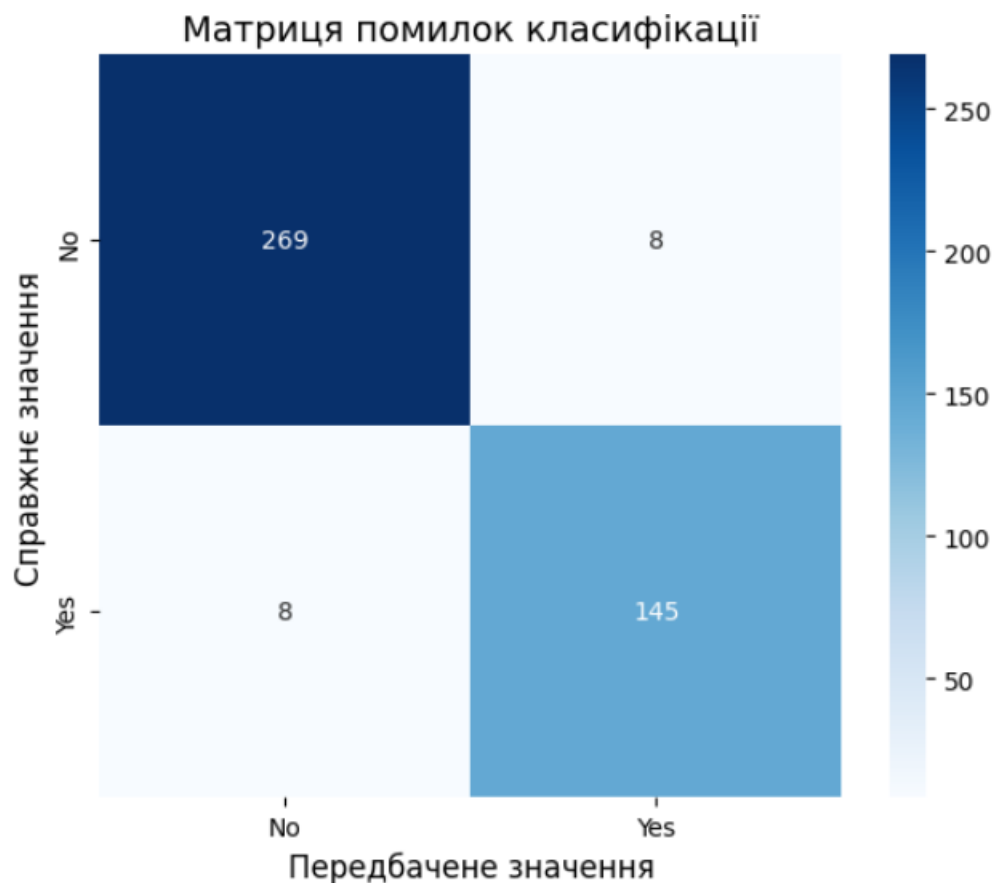


Рисунок 3.12 - Матриця помилок Stacking Ensemble

Отримані результати дозволяють зробити обґрунтований висновок, що модель StackingClassifier є ефективним інструментом для вирішення задачі прогнозування ризику виникнення хвороби Альцгеймера у цьому дослідженні.

Висока точність і надійність роблять цю модель доречною для використання у клінічних інформаційних системах, зокрема в системах підтримки прийняття рішень, пов'язаних із попереднім скринінгом пацієнтів на предмет розвитку нейродегенеративних захворювань.

3.5 Висновки до розділу 3

У результаті порівняльного аналізу класифікаційних моделей для прогнозування ризику виникнення захворювання Альцгеймера можна зробити висновок, що найбільш ефективною виявилася модель StackingClassifier, яка продемонструвала найвищі значення всіх ключових метрик (Accuracy, Precision, Recall, F1-Score, AUC) та мінімальну кількість хибнонегативних передбачень (рис. 3.13).



Рисунок 3.13 – Порівняння отриманих метрик кожної з моделей

Варто зазначити, що порівняно з моделлю дерева рішень, Random Forest трохи поступається за AUC та recall, але має меншу кількість хибнопозитивних передбачень, що може бути критично важливим у клінічному контексті. Також на основі наведених метрик можна дійти висновку, що модель SVC демонструє обмежену ефективність для прогнозування ризику виникнення хвороби Альцгеймера. Її точність, чутливість, і особливо AUC є істотно нижчими у порівнянні з іншими моделями.

Отже, StackingClassifier є оптимальним вибором для вирішення задачі прогнозування ризику виникнення хвороби Альцгеймера, завдяки найвищим показникам точності, чутливості та дискримінативної здатності. Хоча різниця між моделями StackingClassifier та XGBoost є незначною з точки зору кількісних показників, однак StackingClassifier має трохи меншу кількість хибнопозитивних і хибнонегативних випадків, узагальнення закономірностей у даних та більш збалансованого розпізнавання обох класів — як пацієнтів із ризиком захворювання, так і здорових осіб, що є критично важливим у медичних задачах, де помилка будь-якого типу може мати серйозні наслідки для діагностики та подальшого лікування. Попри більшу архітектурну та обчислювальну складність StackingClassifier, яка передбачає поєднання кількох моделей в одну мета-модель, її здатність до балансування помилок обох типів може мати переваги у чутливих медичних задачах, де кожне передбачення може вплинути на своєчасність діагностики. Високі результати StackingClassifier зумовлені синергетичним ефектом поєднання різних алгоритмів, які компенсують слабкі сторони один одного. Для подальших досліджень доцільно розглянути оптимізацію архітектури стекінгової моделі, використання більшої кількості медичних показників, інтеграції часових змінних для прогнозування динаміки захворювання, а також оцінку узагальнення на незалежних валідаційних вибірках, зокрема з інших клінік для перевірки їхньої здатності до узагальнення у клінічній практиці.

РОЗДІЛ 4 ФУНКЦІОНАЛЬНО-ВАРТІСНИЙ АНАЛІЗ ПРОГРАМНОГО ПРОДУКТУ

У цьому розділі здійснено оцінювання ключових характеристик програмного продукту, створеного для прогнозування ризику виникнення захворювання Альцгеймера із використанням методів машинного навчання. Метою даного етапу є визначення ефективності реалізації, обґрунтування вибору технічних рішень і забезпечення раціонального співвідношення між вартістю розробки, продуктивністю системи та її функціональними можливостями.

Розглядалися різні варіанти реалізації програмного забезпечення з урахуванням вимог до обробки медичних даних, сумісності з апаратним забезпеченням, інтерфейсної зручності та адаптивності до подальшого масштабування. Для обґрунтування оптимального варіанту реалізації було застосовано апарат функціонально-вартісного аналізу (ФВА).

Функціонально-вартісний аналіз є методологією, що дозволяє оцінити реальну вартість продукту або послуги незалежно від організаційної структури компанії. Він спрямований на виявлення резервів зниження витрат завдяки підвищенню ефективності функціональних компонентів, оптимізації ресурсів та забезпеченню кращого співвідношення між споживчою цінністю продукту та витратами на його розробку й впровадження.

У процесі проведення ФВА була використана технічна, економічна та конструкторська інформація.

Алгоритм функціонально-вартісного аналізу включає в себе визначення послідовності етапів розробки програмного забезпечення, визначення повних витрат (річних) та кількості робочих часів, визначення джерел витрат та підсумковий розрахунок вартості програмного продукту.

4.1 Постановка задачі проектування

У роботі застосовується метод ФВА для проведення техніко-економічного аналізу розробки системи для прогнозування ризику виникнення захворювання Альцгеймера. Оскільки рішення стосовно проектування та реалізації компонентів, що розробляється, впливають на всю систему, кожна окрема підсистема має її задовольняти. Тому фактичний аналіз представляє собою аналіз функцій програмного продукту, призначеного для збору, обробки та проведення аналізу даних по захворюванню Альцгеймера.

Технічні вимоги до програмного продукту є наступні:

- функціонування на персональних комп'ютерах із стандартним набором апаратних та програмних компонентів;
- зручність та інтуїтивна зрозумілість для кінцевого користувача;
- швидкість обробки даних та доступ до інформації в реальному часі;
- можливість зручного масштабування та обслуговування;
- мінімальні витрати на впровадження програмного продукту.

4.2 Обґрунтування функцій програмного продукту

Головна функція F_0 – розробка програмного продукту, за допомогою якого виконується порівняльний аналіз різних характеристик роботи обраних для дослідження методів для прогнозування ризику виникнення захворювання Альцгеймера, що безпосередньо впливають на стійкість підприємства. Беручи за основу цю функцію, можна виділити наступні підфункції:

- F_1 – вибір мови програмування;
- F_2 – вибір обчислювальних потужностей;
- F_3 – вибір бібліотеки.

Кожна з цих функцій має декілька варіантів реалізації:

- функція F_1 :
 - а) Python;
 - б) C++.
- функція F_2 :
 - а) PyCharm;
 - б) Visual Studio Code (VS Code);
 - в) Jupyter Lab.
- функція F_3 :
 - а) Застосування вбудованих функцій;
 - б) Створення своїх функцій.

Варіанти реалізації основних функцій наведені у морфологічній карті системи (рис. 4.1).

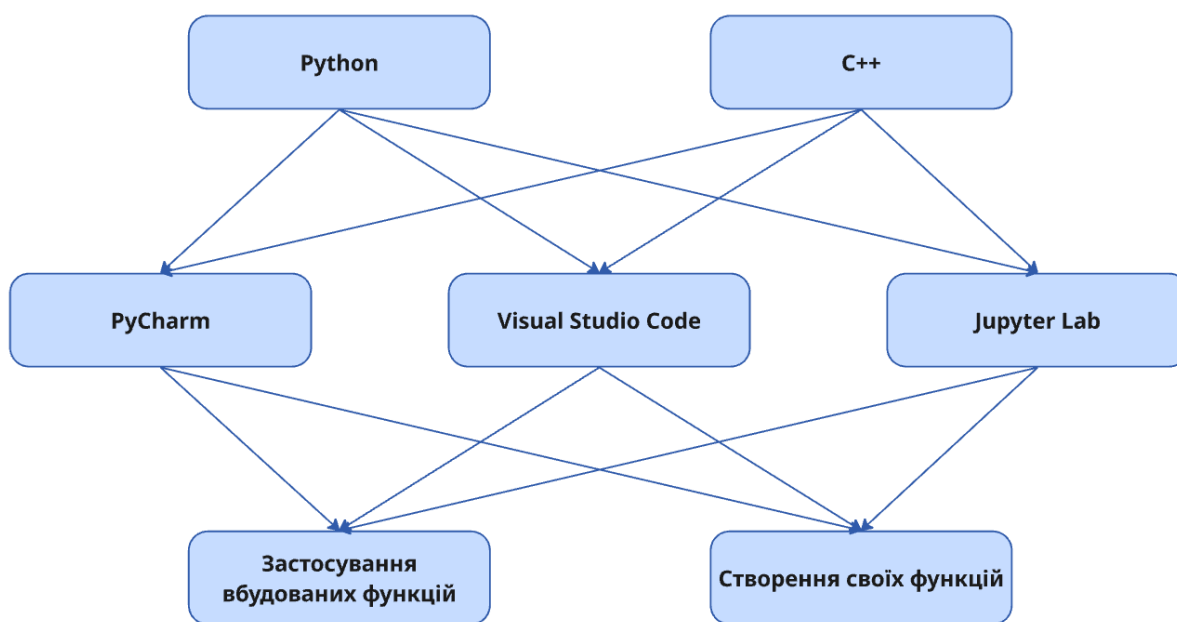


Рисунок 4.1 – Морфологічна карта

Морфологічна карта відображає множину всіх можливих варіантів основних функцій. У таблиці 4.1 наведено позитивно-негативну матрицю.

Таблиця 4.1 – Позитивно-негативна матриця

Функції	Варіанти реалізації	Переваги	Недоліки
F_1	А	Розвинена екосистема бібліотек і фреймворків для машинного навчання; простота у використанні;	Повільніша швидкість виконання у порівнянні з компільованими мовами; менший контроль над ресурсами.
	Б	Висока продуктивність і ефективне управління пам'яттю; придатність для ресурсомістких обчислень.	Складність у реалізації алгоритмів машинного навчання; відсутність широкої підтримки ML-бібліотек.
F_1	А	Зручне середовище розробки з інтеграцією бібліотек машинного навчання; покращена підтримка рефакторингу, автодоповнення та тестування.	Обмежена підтримка інтерактивних експериментів; ресурсомістке середовище.
	Б	Легкість, гнучкість, велика кількість ML-плагінів; підтримка швидкого налаштування середовища для проєктів з Python.	Вимагає налаштування розширень для повноцінної підтримки ML; менш зручний для покрокового відлагодження.
	В	Підходить для аналізу даних, експериментів з ML-моделями та візуалізації результатів; підтримка інтерактивного навчання моделей.	Використовує більше ресурсів та може бути обмежений для великих обчислювальних завдань.
F_1	А	Прискорення розробки моделей за рахунок використання перевірених та оптимізованих ML-методів	Обмежена гнучкість при потребі реалізувати нові або спеціалізовані алгоритми.
	Б	Повний контроль над реалізацією ML-алгоритмів; можливість адаптації моделей під специфіку даних та задачі.	Потребує значних зусиль для розробки, тестування та оптимізації; вищий ризик помилок у реалізації.

На основі аналізу позитивно-негативної матриці робимо висновок, що при розробці програмного продукту деякі варіанти реалізації функцій варто

відкинути, тому, що вони не відповідають поставленим перед програмним продуктом задачам. Ці варіанти відзначені у морфологічній карті.

Функція F_1 :

Оскільки реалізація програмного продукту базується на активному використанні бібліотек машинного навчання (scikit-learn), важливо забезпечити швидку інтеграцію з наявними фреймворками, підтримку обробки даних та простоту реалізації моделей. Мова програмування C++ (варіант Б) не забезпечує зручної роботи з ML-бібліотеками та значно ускладнює розробку. Перевагу надаємо Python (варіант А), а варіант Б має бути відкинутий через складність та низьку адаптивність до задач машинного навчання.

Функція F_2 :

Для реалізації програмного продукту, орієнтованого на машинне навчання, доцільно використовувати Jupyter Lab (варіант В) завдяки його інтерактивності, зручності візуалізації результатів та підтримці аналітичного програмування. PyCharm (варіант А) також може бути використаний як повнофункціональне середовище розробки. Натомість Visual Studio Code (варіант Б) менш придатний для реалізації задач даного типу через обмежену інтеграцію з інструментами машинного навчання. Отже, розглядаються варіанти А та В, варіант Б відкидається.

Функція F_3 :

Для ефективної реалізації задачі прогнозування доцільно використовувати перевірені реалізації алгоритмів машинного навчання. Створення власних обчислень (варіант Б) потребує багато часу та підвищує ризик помилок, тому перевагу надаємо вбудованим функціям (варіант А). Отже, варіант Б доцільно виключити з розгляду.

Таким чином, будемо розглядати такі варіанти реалізації ПП:

$$F_1: a - F_2: a - F_1: a$$

$$F_1: a - F_2: b - F_1: a$$

Для оцінювання якості розглянутих функцій обрана система параметрів, описана нижче.

4.3 Обґрунтування системи параметрів програмного продукту

На основі даних, розглянутих вище, визначаються основні параметри вибору, які будуть використані для розрахунку коефіцієнта технічного рівня.

Для того, щоб охарактеризувати програмний продукт, будемо використовувати наступні параметри:

- X1 – швидкодія мови програмування;
- X2 – об’єм пам’яті для обчислень та збереження даних;
- X3 – час навчання даних;
- X4 – потенційний об’єм програмного коду.

Гірші, середні і кращі значення параметрів вибираються на основі вимог замовника й умов, що характеризують експлуатацію програмного продукту, як показано у таблиці 4.2.

Таблиця 4.2 – Основні параметри програмного продукту

Назва параметра	Умовні позначення	Одиниці виміру	Значення параметра		
			гірші	середні	кращі
Швидкодія мови програмування	X1	оп/мс	300	450	700
Об’єм пам’яті	X2	Мб	1024	512	256
Час попередньої обробки даних	X3	мс	80	70	60
Потенційний об’єм програмного коду	X4	кількість рядків коду	1000	450	120

За даними таблиці 4.2 будуються графічні характеристики параметрів – рис. 4.2 – рис. 4.5.

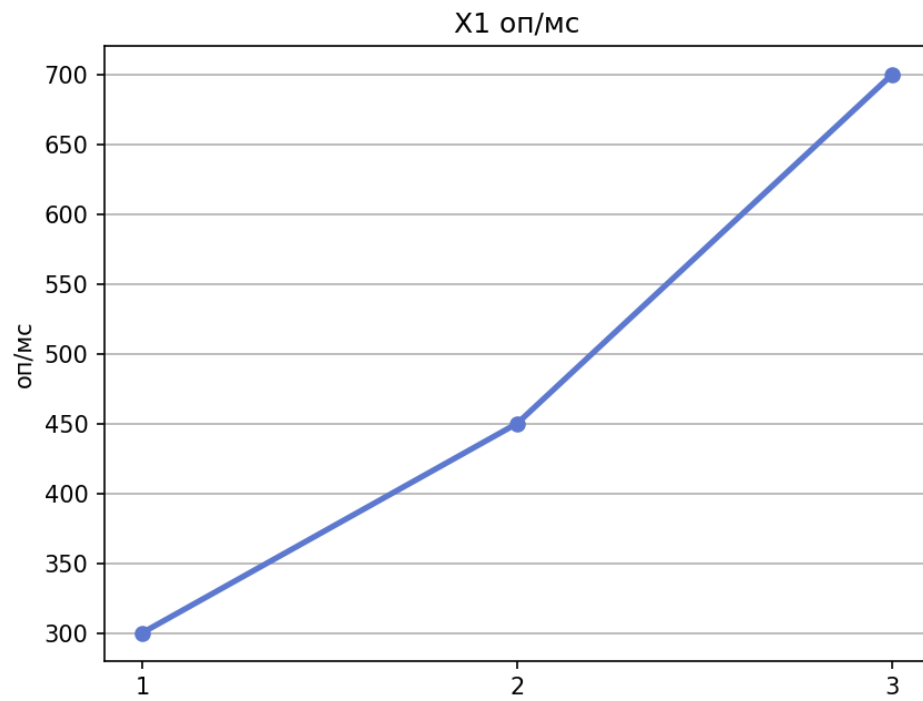


Рисунок 4.2 – X1, швидкодія мови програмування

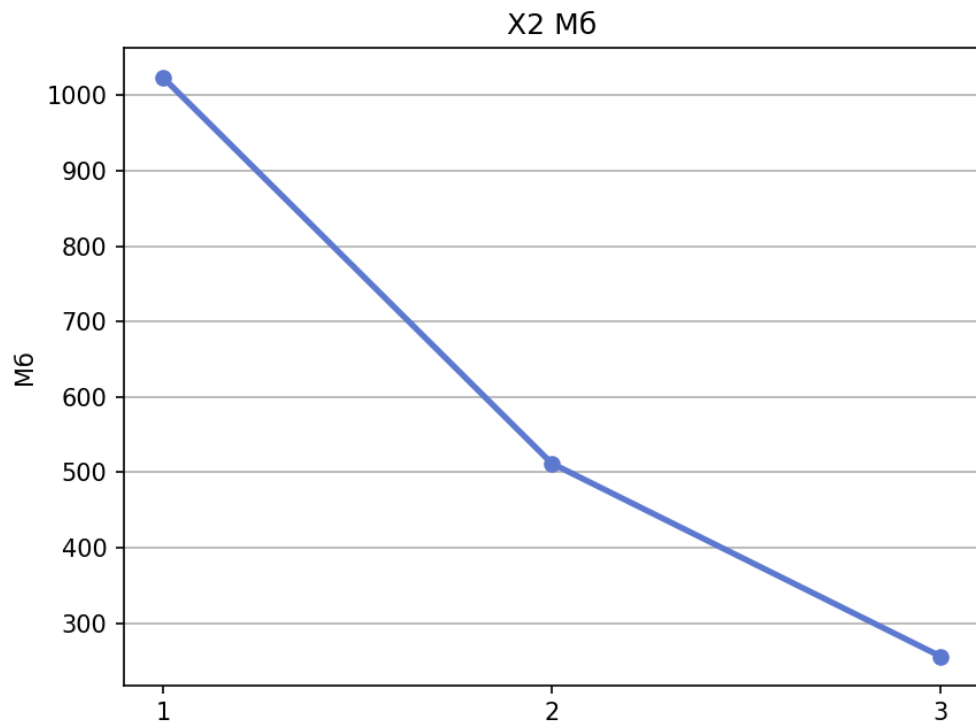


Рисунок 4.3 – X2, об'єм пам'яті

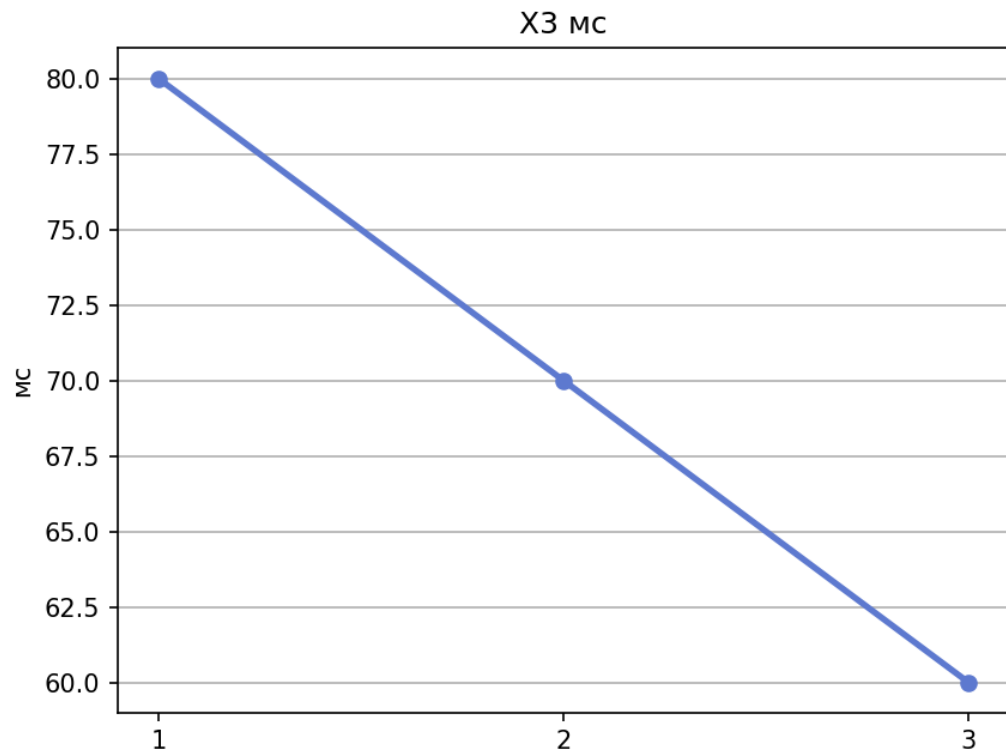


Рисунок 4.4 – X3, час попередньої обробки даних

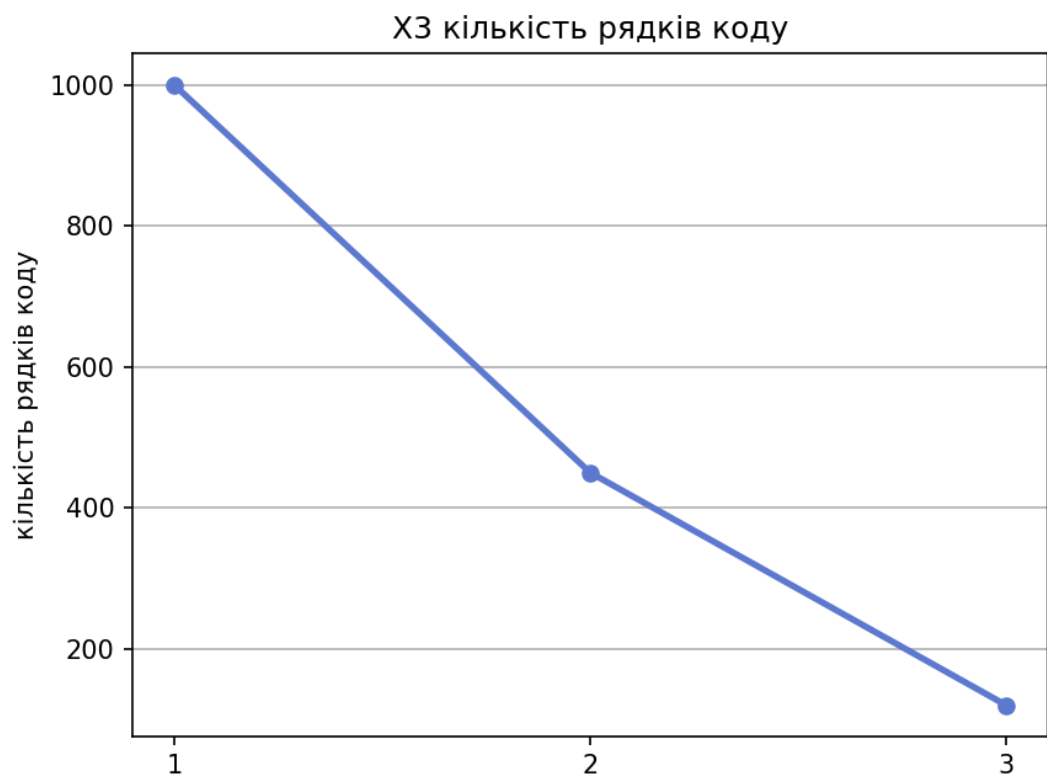


Рисунок 4.5 – X4, потенційний об'єм програмного коду

4.4 Аналіз експертного оцінювання параметрів

Після детального обговорення й аналізу кожний експерт оцінює ступінь важливості кожного параметру для конкретно поставленої цілі – розробка фреймворку, який дає найбільш точні результати при знаходженні параметрів моделей адаптивного прогнозування і обчислення прогнозних значень.

Значимість кожного параметра визначається методом попарного порівняння. Оцінку проводить експертна комісія із 7 людей. Визначення коефіцієнтів значимості передбачає:

- визначення рівня значимості параметра шляхом присвоєння різних рангів;
- перевірку придатності експертних оцінок для подальшого використання;
- визначення оцінки попарного пріоритету параметрів;
- обробку результатів та визначення коефіцієнту значимості.

Результати експертного ранжування наведені у таблиці 4.3.

Таблиця 4.3 – Результати ранжування параметрів

[illegible]

Для перевірки степені достовірності експертних оцінок, визначимо наступні параметри:

а) сума рангів кожного з параметрів і загальна сума рангів:

$$R_i = \sum_{j=1}^N r_{ij} R_{ij} = \frac{Nn(n+1)}{2} = \frac{7 \cdot 4 \cdot 5}{2} = 70 \quad (4.1)$$

де N – число експертів, n – кількість параметрів;

б) середня сума рангів:

$$T = \frac{1}{n} R_{ij} = \frac{70}{4} = 17.5 \quad (4.2)$$

в) відхилення суми рангів кожного параметра від середньої суми рангів:

$$\Delta_i = R_i - T \quad (4.3)$$

Сума відхилень по всіх параметрах повинна дорівнювати 0;

г) загальна сума квадратів відхилення:

$$S = \sum_{i=1}^N \Delta_i^2 = 195 \quad (4.4)$$

Порахуємо коефіцієнт узгодженості:

$$W = \frac{12S}{N^2(n^3 - n)} = \frac{12 \cdot 195}{7^2(4^3 - 4)} \approx 0.795 > W_k = 0.67 \quad (4.5)$$

Ранжування можна вважати достовірним, тому що знайдений коефіцієнт узгодженості перевищує нормативний, котрий дорівнює 0.67.

Скориставшись результатами ранжирування, проведемо попарне порівняння всіх параметрів і результати занесемо у таблицю 4.4.

Таблиця 4.4 – Попарне порівняння параметрів

Параметри	Експерти							Кінцева оцінка	Числове значення
	1	2	3	4	5	6	7		
X1 і X2	<	<	<	<	<	<	<	<	0.5
X1 і X3	<	<	<	>	<	<	<	<	0.5
X1 і X4	<	>	>	>	<	>	>	>	1.5
X2 і X3	>	>	<	>	>	>	>	>	1.5
X2 і X4	>	>	>	>	>	>	>	>	1.5
X3 і X4	>	>	>	>	>	>	>	>	1.5

Числове значення, що визначає ступінь переваги i -го параметра над j -тим, a_{ij} визначається по формулі:

$$a_{ij} = \begin{cases} 1.5 \text{ при } X_i > X_j \\ 1.0 \text{ при } X_i = X_j \\ 0.5 \text{ при } X_i < X_j \end{cases} \quad (4.6)$$

З отриманих числових оцінок переваги складемо матрицю $A = \|a_{ij}\|$. Для кожного параметра зробимо розрахунок вагомості K_{bi} за наступними формулами:

$$K_{bi} = \frac{b_i}{\sum_{i=1}^n b_i} \quad (4.7)$$

$$b_i = \sum_{j=1}^N a_{ij} \quad (4.8)$$

Відносні оцінки розраховуються декілька разів доти, поки наступні значення не будуть незначно відрізнятися від попередніх (менше 2%). На другому і наступних кроках відносні оцінки розраховуються за наступними формулами:

$$K_{bi} = \frac{b'_i}{\sum_{i=1}^n b'_i} \quad (4.9)$$

$$b'_i = \sum_{j=1}^N a_{ij} b_j \quad (4.10)$$

Як видно з таблиці 4.5, різниця значень коефіцієнтів вагомості не перевищує 2%, тому більшої кількості ітерацій не потрібно.

Таблиця 4.5 – Розрахунок вагомості параметрів

Параметри X_i	Параметри X_j				Перша ітер.		Друга ітер.		Третя ітер.	
	X1	X2	X3	X4	b_i	K_{Bi}	b_i^1	K_{Bi}^1	b_i^2	K_{Bi}^2
X1	1	0.5	0.5	1.5	3.5	0.22	12.25	0.20	44.875	0.2
X2	1.5	1	1.5	1.5	5.5	0.34	21.25	0.36	77.875	0.36
X3	1.5	0.5	1	1.5	4.5	0.28	16.25	0.28	59.125	0.28
X4	0.5	0.5	0.5	1	2.5	0.16	9.25	0.16	34.125	0.16
Всього:					16	1	59	1	216	1

4.5 Аналіз рівня якості варіантів реалізації функцій

Визначаємо рівень якості кожного варіанту виконання основних функцій окремо.

Абсолютні значення параметрів X1 (швидкість роботи мови програмування), X2 (об'єм пам'яті) та X3 (час попередньої обробки даних) відповідають технічним вимогам умов функціонування даного ПП.

Абсолютне значення параметра X4 (потенційний об'єм програмного коду) обрано не найгіршим.

Коефіцієнт технічного рівня для кожного варіанта реалізації ПП розраховується так (таблиця 4.6):

$$K_K(j) = \sum_{i=1}^n K_{Bi,j} B_{i,j} \quad (4.11)$$

де n – кількість параметрів; K_{Bi} – коефіцієнт вагомості i -го параметра; B_i – оцінка i -го параметра в балах.

Таблиця 4.6 – Розрахунок показників рівня якості варіантів реалізації основних функцій ПП

Основні функції	Варіант реалізації функції	Параметри	Абсолютне значення параметра	Бальна оцінка параметра	Коефіцієнт вагомості параметра	Коефіцієнт рівня якості
F_1	A	X1	570	25	0.2	5
F_2	A	X2	512	19	0.36	6.84
	B	X3	60	30	0.28	8.4
F_3	A	X4	300	23	0.16	3.68

За даними з таблиці 4.6 за формулою:

$$K_K = K_{Ty}(F_{1k}) + K_{Ty}(F_{2k}) + \dots + K_{Ty}(F_{zk}) \quad (4.12)$$

визначаємо рівень якості кожного з варіантів:

$$K_{K1} = 5 + 6.84 + 3.68 = 15.52$$

$$K_{K2} = 5 + 8.4 + 3.68 = 17.08$$

Як видно з розрахунків, кращим є другий варіант, для якого коефіцієнт технічного рівня має найбільше значення, проте різниця доволі маленька порівняно з першим варіантом.

4.6 Економічний аналіз варіантів розробки ПП

Для визначення вартості розробки ПП спочатку проведемо розрахунок трудомісткості.

Всі варіанти включають в себе два окремих завдання:

1. розробка проекту програмного продукту;
2. розробка програмної оболонки.

Завдання 1 за ступенем новизни відноситься до групи А, завдання 2 – до групи Б. За складністю алгоритми, які використовуються в завданні 1 належать до групи 1; а в завданні 2 – до групи 3.

Для реалізації завдання 1 використовується довідкова інформація, а завдання 2 використовує інформацію у вигляді даних.

Проведемо розрахунок норм часу на розробку та програмування для кожного з завдань.

Загальна трудомісткість обчислюється як:

$$T_o = T_p \cdot K_{\Pi} \cdot K_{СК} \cdot K_M \cdot K_{СТ} \cdot K_{СТ.М} \quad (4.13)$$

де T_p – трудомісткість розробки ПП; K_{Π} – поправочний коефіцієнт; $K_{СК}$ – коефіцієнт на складність вхідної інформації; K_M – коефіцієнт рівня мови програмування; $K_{СТ}$ – коефіцієнт використання стандартних модулів і прикладних програм; $K_{СТ.М}$ – коефіцієнт стандартного математичного забезпечення.

Для першого завдання, виходячи із норм часу для завдань розрахункового характеру степеню новизни А та групи складності алгоритму 1, трудомісткість дорівнює: $T_p = 37$ людино-днів. Поправочний коефіцієнт, який враховує вид нормативно-довідкової інформації для першого завдання: $K_{\Pi} = 1.8$. Поправочний коефіцієнт, який враховує складність контролю вхідної та вихідної інформації для всіх семи завдань рівний 1: $K_{СК} = 1$. Оскільки при розробці першого завдання використовуються стандартні модулі, врахуємо це за допомогою коефіцієнта $K_{СТ} = 0.9$. Тоді загальна трудомісткість програмування першого завдання дорівнює:

$$T_1 = 37 \cdot 1.8 \cdot 0.9 = 59,94 \text{ людино-днів.}$$

Проведемо аналогічні розрахунки для подальших завдань. Для другого завдання (використовується алгоритм третьої групи складності, степінь новизни Б), тобто $T_p = 29$ людино-днів, $K_{\Pi} = 0.9$, $K_{СК} = 1$, $K_{СТ} = 0.8$:

$$T_2 = 29 \cdot 0.9 \cdot 0.8 = 20.88 \text{ людино-днів.}$$

Складаємо трудомісткість відповідних завдань для кожного з обраних варіантів реалізації програми, щоб отримати їх трудомісткість:

$$T_I = (59,94 + 20.88 + 4.8 + 20.88) \cdot 8 = 852 \text{ людино-годин.}$$

$$T_{II} = (59,94 + 20.88 + 6.91 + 20.88) \cdot 8 = 868,88 \text{ людино-годин.}$$

Найбільш високу трудомісткість має варіант II. В розробці беруть участь два програмісти з окладом 20000 грн., один аналітик в області даних з окладом 23000 грн.. Визначимо середню зарплату за годину за формулою:

$$C_q = \frac{M}{T_m \cdot t} \text{ грн.} \quad (4.14)$$

де M – місячний оклад працівників; T_m – кількість робочих днів тижень; t – кількість робочих годин в день.

$$C_q = \frac{20000 + 20000 + 23000}{3 \cdot 21 \cdot 8} = 125 \text{ грн.} \quad (4.15)$$

Тоді, розрахуємо заробітну плату за формулою:

$$C_{зп} = C_q \cdot T_i \cdot K_d \quad (4.16)$$

де $C_{\text{ч}}$ – величина погодинної оплати праці програміста; T_i – трудомісткість відповідного завдання; $K_{\text{д}}$ – норматив, який враховує додаткову заробітну плату.

Зарплата розробників за варіантами становить:

$$\text{I. } C_{\text{зп}} = 125 \cdot 852 \cdot 1.2 = 127800 \text{ грн.}$$

$$\text{II. } C_{\text{зп}} = 125 \cdot 868.88 \cdot 1.2 = 130332 \text{ грн.}$$

Відрахування на єдиний соціальний внесок становить 22%:

$$\text{I. } C_{\text{від}} = C_{\text{зп}} \cdot 0.22 = 127800 \cdot 0.22 = 28116 \text{ грн.}$$

$$\text{II. } C_{\text{від}} = C_{\text{зп}} \cdot 0.22 = 130332 \cdot 0.22 = 28673.04 \text{ грн.}$$

Тепер визначимо витрати на оплату однієї машино-години. (СМ).

Оскільки одна ЕОМ обслуговує одного програміста з окладом 20000 грн., з коефіцієнтом зайнятості 0,2 то для однієї машини отримаємо:

$$C_{\text{Г}} = 12 \cdot M \cdot K_3 = 12 \cdot 20000 \cdot 0.2 = 48000 \text{ грн.}$$

З урахуванням додаткової заробітної плати:

$$C_{\text{зп}} = C_{\text{Г}} \cdot (1 + K_3) = 48000 \cdot (1 + 0.2) = 57600 \text{ грн.}$$

Відрахування на соціальний внесок:

$$C_{\text{від}} = C_{\text{зп}} \cdot 0.22 = 57600 \cdot 0.22 = 11520 \text{ грн.}$$

Амортизаційні відрахування розраховуємо при амортизації 25% та вартості ЕОМ – 10000 грн

$$C_A = K_{TM} \cdot K_A \cdot C_{PP} = 1.4 \cdot 0.25 \cdot 10000 = 3500 \text{ грн.},$$

де K_{TM} – коефіцієнт, який враховує витрати на транспортування та монтаж приладу у користувача; K_A – річна норма амортизації; C_{PP} – договірна ціна приладу.

Витрати на ремонт та профілактику розраховуємо як:

$$C_P = K_{TM} \cdot C_{PP} \cdot K_P = 1.4 \cdot 10000 \cdot 0.08 = 1120 \text{ грн.},$$

де K_P – відсоток витрат на поточні ремонти.

Ефективний годинний фонд часу ПК за рік розраховуємо за формулою:

$$\begin{aligned} T_{EF} &= (D_K - D_B - D_C - D_P) \cdot t_3 \cdot K_B = (365 - 104 - 12 - 16) \cdot 8 \cdot 0.30 = \\ &= 559,2 \text{ години}, \end{aligned}$$

де D_K – календарна кількість днів у році; D_B , D_C – відповідно кількість вихідних та святкових днів; D_P – кількість днів планових ремонтів устаткування; t – кількість робочих годин в день; K_B – коефіцієнт використання приладу у часі протягом зміни.

Витрати на оплату електроенергії розраховуємо за формулою:

$$C_{EL} = T_{EF} \cdot N_C \cdot K_3 \cdot C_{EH} = 559,2 \cdot 0,4 \cdot 0,3 \cdot 9,43 = 632.79 \text{ грн.},$$

де N_C – середньо-споживча потужність приладу; K_3 – коефіцієнтом зайнятості приладу; C_{EH} – тариф за 1 КВт-годин електроенергії.

Накладні витрати розраховуємо за формулою:

$$C_H = C_{\text{ПР}} \cdot 0.67 = 10000 \cdot 0.67 = 6700 \text{ грн.}$$

Тоді, річні експлуатаційні витрати будуть:

$$C_{\text{ЕКС}} = C_{\text{ЗП}} + C_{\text{ВІД}} + C_A + C_P + C_{\text{ЕЛ}} + C_H \quad (4.17)$$

$$C_{\text{ЕКС}} = 57600 + 11520 + 3500 + 1120 + 632.79 + 6700 = 81072.79 \text{ грн.}$$

Собівартість однієї машино-години ЕОМ дорівнюватиме:

$$C_{\text{М-Г}} = \frac{C_{\text{ЕКС}}}{T_{\text{ЕФ}}} = \frac{81072.79}{559.2} = 144.98 \text{ грн/год.}$$

Оскільки в даному випадку всі роботи, які пов'язані з розробкою програмного продукту ведуться на ЕОМ, витрати на оплату машинного часу, в залежності від обраного варіанта реалізації, складає:

$$C_M = C_{\text{М-Г}} \cdot T \quad (4.18)$$

$$\text{I. } C_M = 144.98 \cdot 852 = 123522.96 \text{ грн.}$$

$$\text{II. } C_M = 144.98 \cdot 868.88 = 125970.222 \text{ грн.}$$

Накладні витрати складають 67% від заробітної плати:

$$C_H = C_{\text{ЗП}} \cdot 0.67 \quad (4.19)$$

$$\text{I. } C_H = 127800 \cdot 0.67 = 85626 \text{ грн.}$$

$$\text{II. } C_H = 130332 \cdot 0.67 = 87322.44 \text{ грн.}$$

Отже, вартість розробки ПП за варіантами становить:

$$C_{\text{ПП}} = C_{\text{ЗП}} + C_{\text{ВІД}} + C_{\text{М}} + C_{\text{Н}} \quad (4.20)$$

$$\text{I. } C_{\text{ПП}} = 127800 + 28116 + 123522.96 + 85626 = 365064.96 \text{ грн.}$$

$$\text{II. } C_{\text{ПП}} = 130332 + 28673.04 + 125970.222 + 87322.44 = 372297.702 \text{ грн.}$$

4.7 Вибір кращого варіанту ПП техніко-економічного рівня

Розрахуємо коефіцієнт техніко-економічного рівня за формулою:

$$K_{KEPj} = \frac{K_{Kj}}{C_{\Phi j}} \quad (4.21)$$

$$K_{KEP1} = \frac{15.52}{365064.96} = 4.2513 \cdot 10^{-5}$$

$$K_{KEP2} = \frac{17.08}{372297.702} = 4.58773 \cdot 10^{-5}$$

Як бачимо, найбільш ефективним є перший варіант реалізації програми з коефіцієнтом техніко-економічного рівня $K_{KEP2} = 4.58773 \cdot 10^{-5}$

Після виконання функціонально-вартісного аналізу програмного комплексу що розроблюється, можна зробити висновок, що з альтернатив, що залишились після першого відбору двох варіантів виконання програмного комплексу оптимальним є перший варіант реалізації програмного продукту. У нього виявився найкращий показник техніко-економічного рівня якості $K_{KEP2} = 4.58773 \cdot 10^{-5}$

Цей варіант реалізації програмного продукту має такі параметри:

- вибір мови програмування – Python;
- вибір обчислювальних потужностей – Jupyter Lab;
- вибір бібліотеки – Застосування вбудованих функцій.

Даний варіант виконання програмного комплексу дає користувачу зручний інтерфейс, швидку реалізацію програми та доступний функціонал для роботи.

4.8 Висновки до розділу 4

У цьому розділі було здійснено повний функціонально-вартісний аналіз програмного продукту, призначеного для прогнозування ризику виникнення захворювання Альцгеймера з використанням методів машинного навчання. Також визначено та проведено оцінку основних функцій програмного продукту. Оцінено ключові параметри, що характеризують програмний продукт, такі як швидкодія, обсяг пам'яті, час обробки даних та потенційний обсяг коду. На підставі експертного оцінювання сформовано оптимальну конфігурацію реалізації, яка забезпечує найкраще співвідношення між функціональністю, витратами ресурсів та ефективністю роботи. Зроблені висновки стали основою для обрання варіанту розробки програмного продукту.

ВИСНОВКИ

У рамках виконання цієї бакалаврської дипломної роботи було досліджено проблему застосування методів машинного навчання для прогнозування ризику виникнення захворювання Альцгеймера. Метою дослідження було визначення ефективних підходів до моделювання прогнозів на основі медичних даних, що дозволяють своєчасно виявляти осіб із підвищеним ризиком розвитку цієї нейродегенеративної патології. Реалізація поставлених завдань охоплювала як теоретичні, так і практичні аспекти машинного навчання, включаючи аналіз предметної області, математичне обґрунтування моделей, практичну реалізацію обраних алгоритмів і функціонально-вартісну оцінку програмного продукту.

У першому розділі дипломної роботи було здійснено ґрунтовне дослідження предметної області, що охоплює сучасні знання про хворобу Альцгеймера, її соціально-економічні наслідки, етіологію, фактори ризику та сучасні діагностичні підходи.

У другому розділі було представлено теоретичні засади машинного навчання, описано основні його типи та алгоритми, що мають потенціал у задачах прогнозування медичних станів. Розглянуто як базові моделі (Decision Tree, Support Vector Machine), так і ансамблеві методи, зокрема Random Forest, XGBoost та Stacking. Описано математичні основи побудови моделей, переваги й недоліки, а також придатність для прогнозування ризику виникнення захворювання Альцгеймера. Додатково було сформовано обґрунтування вибору найбільш релевантних та оптимальних моделей для поставленої задачі.

У третьому розділі було проведено практичну реалізацію обраних моделей на основі відповідного датасету за опомогою мови програмування Python у середовищі Jupyter Notebook. Виконано попередню обробку даних, навчання моделей, здійснено вибір гіперпараметрів та проведено оцінку

результатів за допомогою стандартних метрик класифікації, таких як Accuracy, Precision, Recall, F1-Score та ROC AUC. Аналіз результатів показав, що алгоритм XGBoost та ансамблева модель Stacking продемонстрували найвищі показники точності та стабільності прогнозу. Таким чином, поставлену дослідницьку мету було досягнуто та побудовано ефективну систему, що дозволяє з високою точністю визначати ризики розвитку хвороби Альцгеймера.

У четвертому розділі проведено функціонально-вартісний аналіз розробленого програмного продукту, який підтвердив значущість вибору оптимальних рішень при розробці програмних продуктів, які максимально відповідають критеріям продуктивності та економічної доцільності. У результаті було сформовано основні параметри та визначено найбільш доцільні стратегії, що сприяють розробці продуктивних і ресурсоефективних рішень у заданій сфері.

Отримані результати повністю виправдали початкові гіпотези та очікування. Усі обрані моделі машинного навчання успішно продемонстрували здатність до точного прогнозування ризику захворювання Альцгеймера. Зокрема, найбільш продуктивні алгоритми досягли високих показників точності, понад 90%. Проведений аналіз підтвердив, що застосування моделей машинного навчання у поєднанні з правильною структурою даних може значно покращити діагностичний процес та підвищити ефективність раннього виявлення ризикових станів.

У підсумку, дана робота демонструє значний потенціал застосування методів машинного навчання у сфері когнітивної діагностики, зокрема у контексті захворювання Альцгеймера. Вона слугує надійним підґрунтям для подальших досліджень у цьому напрямку, а також може бути використана для побудови практичних інструментів підтримки прийняття рішень у медицині. Результати дослідження підтверджують, що алгоритми машинного навчання, можуть відігравати ключову роль у боротьбі з нейродегенеративними захворюваннями, відкриваючи нові горизонти для медицини майбутнього.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Kavitha C., Vinodhini Mani, Srividhya S. R., Osamah Ibrahim Khalaf, Carlos Andrés Tavera Romero. Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. *Frontiers in Public Health*. 2022, Vol. 10, P. 1-13.
2. Thomas W. Rowe, Ioanna K. Katzourou, Joshua O. Stevenson-Hoare, Matthew R. Bracher-Smith, Dobril K. Ivanov, Valentina Escott-Price. Machine learning for the life-time risk prediction of Alzheimer's disease: a systematic review. Oxford University Press on behalf of the Guarantors of Brain. 2021, Vol. 14, P. 1-17.
3. Pooja Rani, Rohit Lamba, Ravi Kumar Sachdeva, Karan Kumar, Celestine Iwendi. A machine learning model for Alzheimer's disease. *IET Cyber-Physical Systems: Theory & Applications*. 2024, Vol. 9, Issue 2, P. 125-134.
4. Johanna Wallenstena, Caroline Wachtler, Nenad Bogdanovic, Anna Olofssone, Miia Kivipelto, Linus Jönssoni, Predrag Petrovic, Axel C. Carlsson. Machine learning to detect Alzheimer's disease with data on drugs and diagnoses. *Journal of Prevention of Alzheimer's Disease*. 2025, 12(5): 100115, P. 1-8.
5. World Health Organization. Dementia. URL: <https://www.who.int/news-room/fact-sheets/detail/dementia>.
6. MozOK. Хвороба Альцгеймера і спадковість. URL: <https://mozok.ua/dementia/article/3191-hvoroba-altcgejmera--spadkovst>.
7. Life-House. Поширеність хвороби Альцгеймера в світі та Україні. URL: <https://life-house.ua/blog/hvoroba-altcgejmera-statistika/>.
8. Simiao Chen, Zhong Cao, Arindam Nandi, Nathaniel Counts, Lirui Jiao, Klaus Prettnner, Michael Kuhn, Benjamin Seligman, Daniel Tortorice, Daniel Vigo, Chen Wang, David E Bloom. The global macroeconomic burden of Alzheimer's disease and other dementias: estimates and projections for 152

- countries or territories. *The Lancet Global Health*. 2024, Vol. 12, P. e1534-e1543.
9. Risk factors and risk reduction. *Alzheimer's Disease International*. URL: <https://www.alzint.org/about/risk-factors-risk-reduction/>.
 10. Healthline, Wendy Leonard. Causes and Risk Factors of Alzheimer's Disease. URL: <https://www.healthline.com/health/alzheimers-disease-risk-factors>.
 11. Alzheimer's Associations. What are the Causes and Risk Factors of Alzheimer's and Other Dementias?. URL: <https://www.alz.org/alzheimers-dementia/what-is-alzheimers/causes-and-risk-factors>.
 12. National Library Of Medicine. Alzheimer's Disease: Etiology, Neuropathology and Pathogenesis. URL: <https://www.ncbi.nlm.nih.gov/books/NBK566126/>.
 13. Healthline. What are causes and risk factors for Alzheimer's disease?. URL: <https://www.healthline.com/health/alzheimers-disease-risk-factors>.
 14. Alzheimer's Associations. Medical Tests for Diagnosing Alzheimer's. URL: https://www.alz.org/alzheimers-dementia/diagnosis/medical_tests.
 15. Mayo Clinic. Alzheimer's disease. URL: <https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/diagnosis-treatment/drc-20350453>.
 16. Mayo Clinic. Diagnosing Alzheimer's: How Alzheimer's is diagnosed. URL: <https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers/art-20048075>.
 17. Alzheimer's Society. Risk factors for Alzheimer's disease. URL: <https://www.alzheimers.org.uk/about-dementia/types-dementia/who-gets-alzheimers-disease>.
 18. Muhammad Shahbaz, Shahzad Ali, Aziz Guergachi, Aneeta Niazi, Amina Umer. Classification of Alzheimer's Disease using Machine Learning Techniques. Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA). 2019, P. 296-303.

19. Sergio Grueso, Raquel Viejo-Sobera. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. *Open Access*. 2021, 13:162, P. 1-29.
20. Ji Hwan Park, Han Eol Cho, Jong Hun Kim, Melanie M. Wall, Yaakov Stern, Hyunsun Lim, Shinjae Yoo, Hyoung Seop Kim, Jiook Cha. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *Npj Digital Medicine*. 2020, 3:46, P. 1-7.
21. IBM. What is a decision tree?. URL: <https://www.ibm.com/think/topics/decision-trees>.
22. Kd Nuggets. Decision Tree Algorithm, Explained. URL: <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>.
23. IBM. What is random forest?. URL: <https://www.ibm.com/think/topics/random-forest>.
24. Medium. Random Forest Classification and it's Mathematical Implementation. URL: <https://medium.com/@priyankaparashar54/random-forest-classification-and-its-mathematical-implementation-f5e7e6878ca>.
25. IBM. What is XGBoost?. URL: <https://www.ibm.com/think/topics/xgboost>.
26. Geeksforgeeks. XGBoost. URL: <https://www.geeksforgeeks.org/xgboost/>.
27. IBM. What are support vector machines (SVMs)?. URL: <https://www.ibm.com/think/topics/support-vector-machine>.
28. Geeksforgeeks. Support Vector Machine (SVM) Algorithm. URL: <https://www.geeksforgeeks.org/support-vector-machine-algorithm/>.
29. Medium. Stacking to Improve Model Performance: A Comprehensive Guide on Ensemble Learning in Python. URL: https://medium.com/@brijesh_soni/stacking-to-improve-model-performance-a-comprehensive-guide-on-ensemble-learning-in-python-9ed53c93ce28.
30. Geeksforgeeks. Stacking in Machine Learning. URL: <https://www.geeksforgeeks.org/stacking-in-machine-learning/>.

31. Unite. Ai. Моделі ШІ використовують обробку природної мови для прогнозування ризику розвитку хвороби Альцгеймера. URL: <https://www.unite.ai/uk/ai-models-use-natural-language-processing-to-predict-risk-of-developing-alzheimers/>.
32. Journal of Biomedical Physics and Engineering. Advanced Diagnostic Technique for Alzheimer's Disease using MRI Top-Ranked Volume and Surface-based Features. URL: https://jbpe.sums.ac.ir/article_48580.html.
33. JupyterLab. JupyterLab Documentation. URL: <https://jupyterlab.readthedocs.io/en/latest/>.

ДОДАТКИ

Додаток А. Програмна реалізація алгоритмів

```

# For EDA
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import tensorflow as tf
from sklearn.preprocessing import LabelEncoder, RobustScaler

# For Machine Learning
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
from sklearn.model_selection import train_test_split
from sklearn.ensemble import StackingClassifier
from sklearn.preprocessing import LabelEncoder, StandardScaler
import xgboost as xgb
from sklearn.linear_model import LogisticRegression, BayesianRidge
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, classification_report, confusion_matrix
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import roc_auc_score

df = pd.read_csv('alzheimers_disease_data_2.csv')
df.head()

# Виведемо основні статистичні характеристики числових колонок
датасету

```

```

df.describe().T

df.info()

# Виведення кількості рядків і стовпців у датасеті
print(f"Number of Row : {df.shape[0]}\nNumber of Columns :
{df.shape[1]}")

# Підрахунок кількості пропущених значень у кожному стовпці
pd.DataFrame({'Count':df.shape[0],
              'Null':df.isnull().sum(),
              'Null %':df.isnull().mean() * 100,
              'Cardinality':df.nunique()
            })

# Перевірка на наявність дублікатів
print("Duplicate Rows (%)")
duplicate_count = df.duplicated().mean()
print(duplicate_count)

# Set figure size for better visibility
plt.figure(figsize=(18, 20))

# EDA
# Розподіл даних
# 1. Histogram for Age Distribution
plt.subplot(3, 3, 1)
sns.histplot(df['Age'], bins=30, kde=True, color='green')
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count')

# 3. Histogram for Cognitive Test Score
plt.subplot(3, 3, 2)
sns.histplot(df['EducationLevel'], bins=30, kde=True, color='skyblue')
plt.title('Cognitive Test Score Distribution')
plt.xlabel('Score')

```

```
plt.ylabel('Count')

# 2. Histogram for BMI Distribution
plt.subplot(3, 3, 3)
sns.histplot(df['BMI'], bins=30, kde=True, color='red')
plt.title('BMI Distribution')
plt.xlabel('BMI')
plt.ylabel('Count')

# 5. Count plot for Smoking Status
plt.subplot(3, 3, 4)
sns.countplot(x='Smoking', data=df, palette='Set3')
plt.title('Smoking Status Distribution')
plt.xlabel('Smoking Status')
plt.ylabel('Count')

# 4. Count plot for Gender Distribution
plt.subplot(3, 3, 5)
sns.countplot(x='Gender', data=df, palette='coolwarm')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')

# 6. Count plot for Alzheimer's Diagnosis
plt.subplot(3, 3, 6)
sns.countplot(x="Diagnosis", data=df, palette='Set2')
plt.title("Alzheimer's Diagnosis Distribution")
plt.xlabel('Diagnosis')
plt.ylabel('Count')

# 7. Count plot for Memory Complaints
plt.subplot(3, 3, 7)
sns.countplot(x="MemoryComplaints", data=df, palette='Set2')
plt.title("Memory Complaints")
plt.xlabel('MemoryComplaints')
plt.ylabel('Count')

# 8. Count plot for Alzheimer's Diagnosis
```

```

plt.subplot(3, 3, 8)
sns.countplot(x="Ethnicity", data=df, palette='Set1')
plt.title("Ethnicity")
plt.xlabel('Ethnicity')
plt.ylabel('Count')

# 9. Count plot for Family History of Alzheimer's
plt.subplot(3, 3, 9)
sns.countplot(x="FamilyHistoryAlzheimers", data=df, palette='Set3')
plt.title("Family History of Alzheimer's Distribution")
plt.xlabel('FamilyHistoryAlzheimers')
plt.ylabel('Count')

# Adjust layout for better visualization
plt.tight_layout()

# Show all plots
plt.show()

# Кореляційний аналіз
plt.figure(figsize=(25, 8))
sns.boxplot(data=df.drop(['PatientID'], axis=1))
plt.title('Distribution of the values of all potential standardized
predictors')
plt.xticks(rotation=45)
plt.grid(True)
plt.show()

def encode_categorical_columns(df):
    """
    Кодує категоріальні (object) колонки за допомогою LabelEncoder.
    Повертає змінений датафрейм та словник з енкодерами.
    """
    label_encoders = {}
    for col in df.select_dtypes(include='object').columns:
        encoder = LabelEncoder()

```

```

        df[col] = encoder.fit_transform(df[col])
        label_encoders[col] = encoder
    return df, label_encoders

def scale_numerical_columns(df, exclude_cols):
    """
    Масштабує числові колонки, які не входять до exclude_cols,
    використовуючи StandardScaler.
    Повертає змінений датафрейм.
    """
    numeric_cols = df.select_dtypes(include=['int64',
'float64']).columns
    numeric_cols = numeric_cols.drop(['Diagnosis'])
    cols_to_scale = [col for col in numeric_cols if col not in
exclude_cols]

    scaler = RobustScaler()
    df[cols_to_scale] = scaler.fit_transform(df[cols_to_scale])

    return df, cols_to_scale

# Основний блок обробки
# 1. Кодуємо категоріальні змінні
df, label_encoders = encode_categorical_columns(df)
original_object_cols = list(label_encoders.keys())

# 2. Масштабуємо числові змінні
df, scaled_columns = scale_numerical_columns(df, original_object_cols)

# 3. Виводимо перші кілька рядків масштабованих даних
print("Scaled Numerical Dat:")
df[scaled_columns].head()

plt.figure(figsize=(25, 8))
sns.boxplot(data=df.drop(['PatientID'], axis=1))
plt.title('Distribution of the values of all potential standardized

```

```

predictors')
plt.xticks(rotation=45)
plt.grid(True)
plt.show()

label_encoders = {}
for column in df.select_dtypes(include=['object']).columns:
    label_encoders[column] = LabelEncoder()
    df[column] = label_encoders[column].fit_transform(df[column])

# Побудуємо кореляційну матрицю
plt.figure(figsize=(20,12))
sns.heatmap(df.corr(method='pearson'), annot=True, fmt='.2f',
            cmap='coolwarm', cbar=True)
plt.title("Correlation matrix")
plt.show()

# Порівняння продуктивності моделі машинного навчання
def evaluate_model(model, X_train, X_test, y_train, y_test):
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred, average='weighted',
                                zero_division=0)
    recall = recall_score(y_test, y_pred, average='weighted',
                           zero_division=0)
    f1 = f1_score(y_test, y_pred, average='weighted', zero_division=0)
    auc = roc_auc_score(y_test, y_pred)

    return accuracy, precision, recall, f1, auc

def evaluate_model_map(model):
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    print(f'Accuracy Train Score : {model.score(X_train, y_train)}\n')

```

```

print(f'Accuracy Test Score : {model.score(X_test, y_test)}\n')
print(f'Classification Report: \n{classification_report(y_test,
y_pred)}\n')
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6, 5))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=['No', 'Yes'],
            yticklabels=['No', 'Yes'],
            cbar=True, square=True)

plt.title('Матриця помилок класифікації', fontsize=14)
plt.xlabel('Передбачене значення', fontsize=12)
plt.ylabel('Справжнє значення', fontsize=12)
plt.tight_layout()
plt.show()

X = df.drop(['Diagnosis'], axis=1)
y = df['Diagnosis']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

estimators = [
    ('svm', SVC(probability=True, random_state=42)),
    ("Random Forest", RandomForestClassifier(n_estimators=100,
max_depth=10, min_samples_split=5, min_samples_leaf=1,
max_features='sqrt', random_state=42)),
    ('xgb', XGBClassifier(use_label_encoder=False,
eval_metric='logloss', random_state=42))
]

models = {
    "Decision Tree": DecisionTreeClassifier(max_depth=10,
# limits tree depth
    min_samples_split=2,      # requires at least 10 samples to split
    min_samples_leaf=4,       # each leaf must have at least 5 samples
    max_leaf_nodes=20,        # limit number of leaf nodes
    random_state=42

```

```

),
    "Random Forest": RandomForestClassifier(n_estimators=50,          #
number of trees
    max_depth=None,          # max depth of each tree
    min_samples_split=5,     # min samples to split
    min_samples_leaf=2,      # min samples per leaf
    max_features='sqrt',     # reduce features used per split
    random_state=42
),
    "XGB": xgb.XGBClassifier(objective='multi:softmax', num_class=3,
    # reg_alpha=0.1, # L1 penalty
    # reg_lambda=1.0, # L2 penalty
    colsample_bytree=0.8, learning_rate=0.3, max_depth=5,
n_estimators=100, subsample=1.0,
    random_state=42),

    "SVC": SVC(C=1.0, gamma='scale', kernel='rbf', degree=3),

    "StackingClassifier": StackingClassifier(estimators=estimators,
final_estimator=LogisticRegression(), cv=5, passthrough=False,
n_jobs=-1, stack_method = 'auto'),
}

results = []

for name, model in models.items():
    accuracy, precision, recall, f1, auc = evaluate_model(model,
X_train, X_test, y_train, y_test)
    results.append({
        "Model": name,
        "Accuracy": accuracy,
        "Precision": precision,
        "Recall": recall,
        "F1-Score": f1,
        "AUC": auc
    })

```

```

results_df = pd.DataFrame(results)
results_df

for name, model in models.items():
    print("Model: ", model)
    evaluate_model_map(model)

# Висновок, обрання найоптимальнішої моделі
best_model = results_df.loc[results_df['Accuracy'].idxmax()]
print("\nBest Model (Machine Learning):")
print(best_model)

metrics = ["Accuracy", "Precision", "Recall", "F1-Score", "AUC"]
colors = ["skyblue", "orange", "green", "red", "pink"]

fig, ax = plt.subplots(figsize=(10, 8))

offset = 0.3

for i, metric in enumerate(metrics):
    ax.barh(results_df['Model'] + f" ({metric})", results_df[metric],
            color=colors[i], alpha=0.7, label=metric, height=0.6)

ax.set_xlabel("Score", fontsize=12)
ax.set_title("Model Performance Comparison", fontsize=18)
ax.legend(title="Metrics")

plt.tight_layout()
plt.show()

print(f"\nOptimal Model based on Accuracy: {best_model['Model']} with
Accuracy = {best_model['Accuracy']:.4f}")

```

Додаток Б. Презентація

МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ РИЗИКУ ВИНИКНЕННЯ ЗАХВОРЮВАННЯ АЛЬЦГЕЙМЕРА

Виконала:
студентка IV курсу, групи КА-14
Петренко Карина Ігорівна

Керівник:
Завідувач кафедри ММСА, к.т.н., доцент
Тимошук Оксана Леонідівна

АКТУАЛЬНІСТЬ ДОСЛІДЖЕННЯ

- Зростаюча поширеність захворювання Альцгеймера
- Відсутність ефективного лікування
- Велика кількість доступних медичних даних та розвиток технологій штучного інтелекту
- Соціальна та економічна значущість проблеми
- Раннє виявлення захворювання та покращення якості життя

Об'єкт дослідження: хвороба Альцгеймера; процес діагностування ризику виникнення захворювання Альцгеймера.

Предмет дослідження: математичні методи машинного навчання для прогнозування ризику виникнення захворювання Альцгеймера.

Мета дослідження: розробка та оцінка ефективності моделей машинного навчання для прогнозування ризику виникнення захворювання Альцгеймера на основі медичних та когнітивних ознак.

3

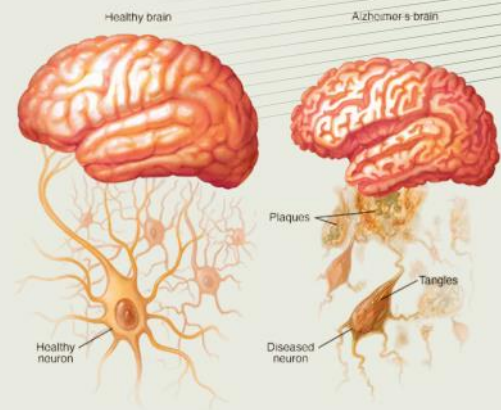
ПОСТАНОВКА ЗАДАЧІ

- Проведення аналізу та вибір алгоритмів машинного навчання оптимальних для прогнозування ризику виникнення захворювання Альцгеймера.
- Підготовка та попередня обробка даних
- Побудова та навчання моделей
- Оцінка та аналіз ефективності роботи досліджуваних методів

4

ПОНЯТТЯ ЗАХВОРЮВАННЯ АЛЬЦГЕЙМЕРА

Захворювання Альцгеймера є однією з найпоширеніших форм деменції, що вражає мільйони людей у світі, особливо людей похилого віку. Це прогресуюче нейродегенеративне захворювання не лише призводить до значного зниження когнітивних функцій, а й має серйозні соціальні та економічні наслідки для пацієнтів, їхніх родин і систем охорони здоров'я загалом, оскільки є хворобою, яка потребує паліативної допомоги.



5

ОБРАНІ МОДЕЛІ ДОСЛІДЖЕННЯ

- Decision Tree
- Random Forest
- XGBoost
- Support Vector Machine
- Stacking Ensemble

6

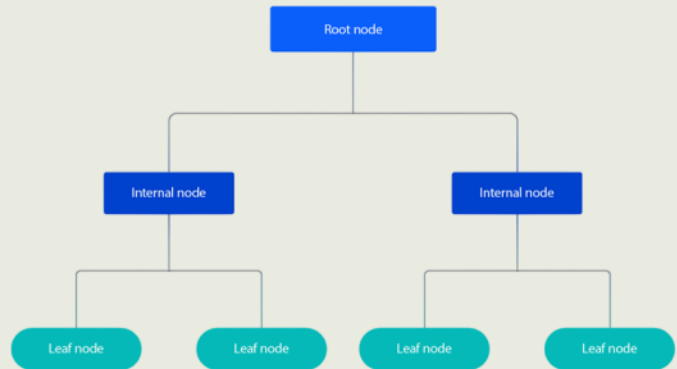
DECISION TREE

Переваги:

- Простота та інтерпретованість
- Універсальність
- Не потребує масштабування ознак
- Обробка нелінійних зв'язків

Недоліки:

- Перенавчання (overfitting)
- Нестабільність
- Схильність до ознак з багатьма категоріями



7

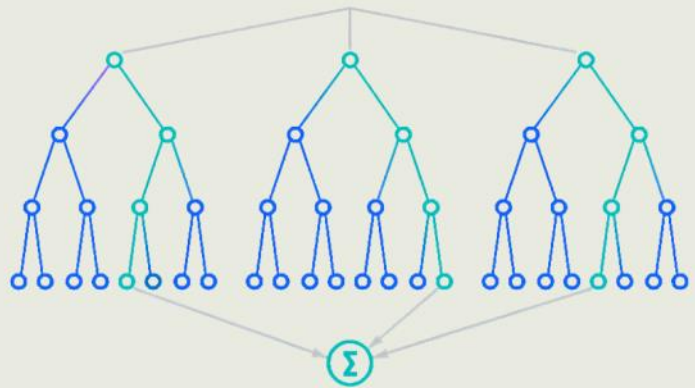
RANDOM FOREST

Переваги:

- Менший ризик перенавчання
- Гнучкість
- Оцінка важливості ознак

Недоліки:

- Повільна обробка
- Високе споживання ресурсів
- Менш інтерпретована



8

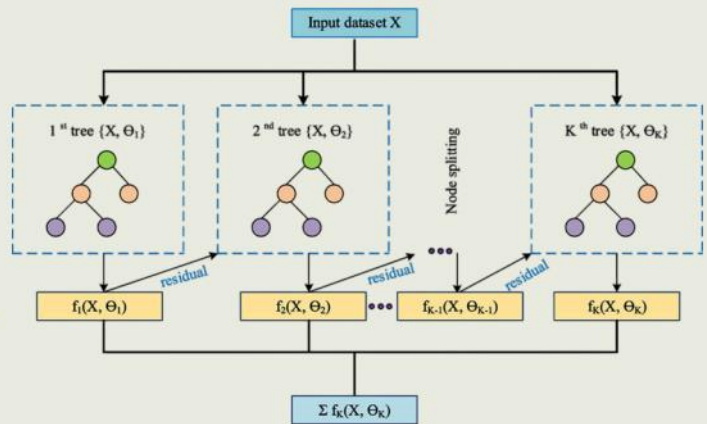
XGBOOST

Переваги:

- Висока продуктивність
- Швидкість навчання
- Універсальність
- Стійкість до перенавчання

Недоліки:

- Необхідність ретельного тюнінгу
- Слабка інтерпретованість



9

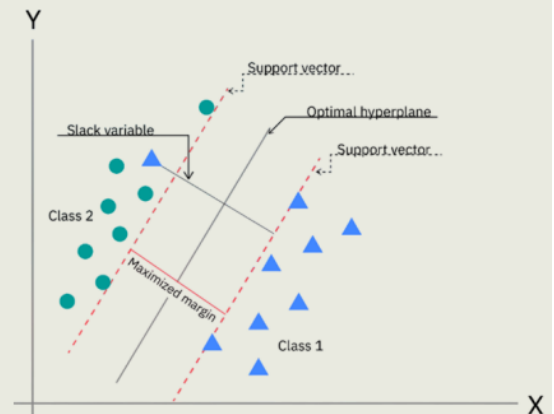
SUPPORT VECTOR MACHINE

Переваги:

- Працює з складними та нелінійними даними
- Стійкість до викидів
- Ефективність у високовимірному просторі
- Схильність до уникнення перенавчання
- Гнучкість через використання ядер

Недоліки:

- Повільне навчання на великих даних
- Складність налаштування
- Чутливість до шуму
- Слабка інтерпретованість
- Високе споживання пам'яті



10

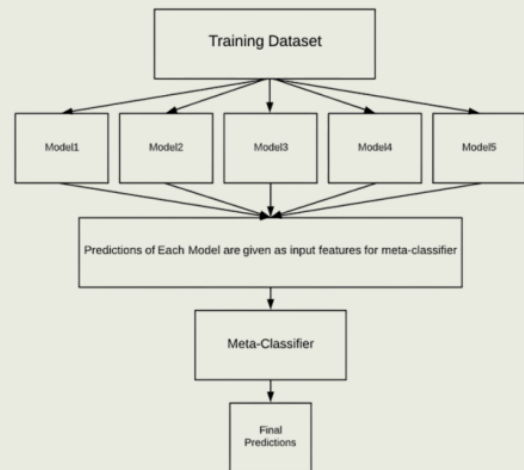
STACKING ENSEMBLE

Переваги:

- Висока точність
- Можливість поєднання різних моделей
- Зменшення ризику перенавчання
- Навчання на помилках базових моделей
- Гнучкість

Недоліки:

- Складність реалізації
- Повільне навчання
- Слабка інтерпретованість
- Потребує більше даних



11

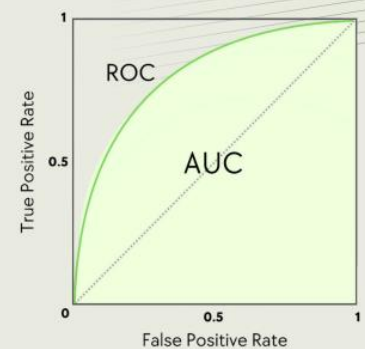
МЕТРИКИ ОЦІНЮВАННЯ

$$\text{Accuracy} = \frac{\text{Кількість правильних передбачень}}{\text{Загальна кількість передбачень}} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{\text{Кількість істинно позитивних передбачень}}{\text{Кількість усіх передбачень позитивного класу}} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{\text{Кількість істинно позитивних передбачень}}{\text{Кількість усіх фактичних позитивних прикладів}} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$



- True Positive - кількість істинно позитивних значень;
- True Negative - кількість істинно негативних значень;

- False Positive - кількість невірно позитивних значень;
- False Negative - кількість невірно негативних значень.

12

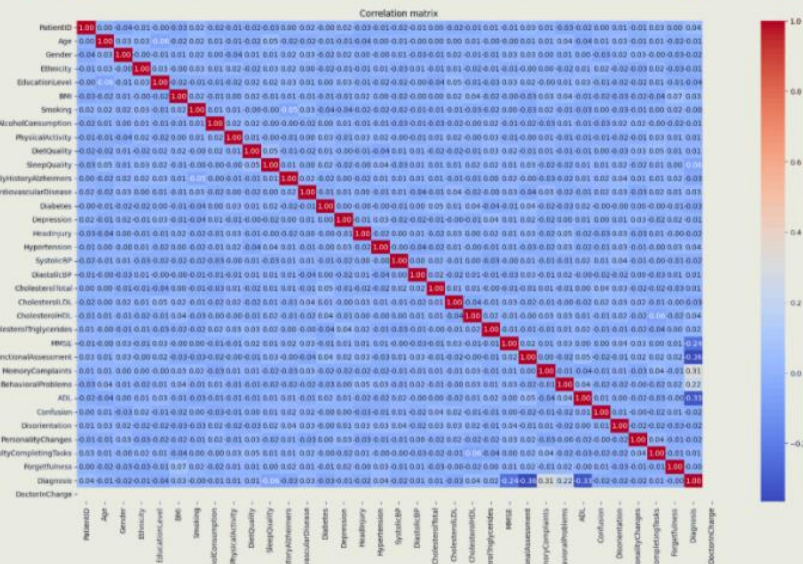
ОПИС НАБОРУ ДАНИХ ALZHEIMER'S DISEASE DATASET

У якості робочого датасету для проведення порівняння ефективності обраних методів у задачі прогнозування ризику виникнення захворювання Альцгеймера було обрано набір даних Alzheimer's Disease Dataset. Цей набір даних містить розширену інформацію про здоров'я 2149 пацієнтів, кожен з яких має унікальний ідентифікатор. Датасет охоплює 35 ознак, що містять комплексну інформацію про пацієнтів, зокрема: демографічні дані, фактори способу життя, медичну історію, клінічні вимірювання, оцінки когнітивного й функціонального стану, а також наявність симптомів та діагноз хвороби Альцгеймера.

	PatientID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity	DietQuality	...	PersonalityChanges	DifficultyCompletingTasks	Forgetfulness	Diagnosis
0	4751	73	0	0	2	22.927749	0	13.297218	6.327112	1.347214	—	0	1	0	0
1	4752	89	0	0	0	26.827681	0	4.542524	7.619885	0.518767	—	0	0	1	0
2	4753	73	0	3	1	17.795882	0	19.555085	7.844988	1.826335	—	0	1	0	0
3	4754	74	1	0	1	33.800817	1	12.209266	8.428001	7.435604	—	0	0	0	0
4	4755	89	0	0	0	20.716974	0	18.454356	6.310461	0.795498	—	1	1	0	0

13

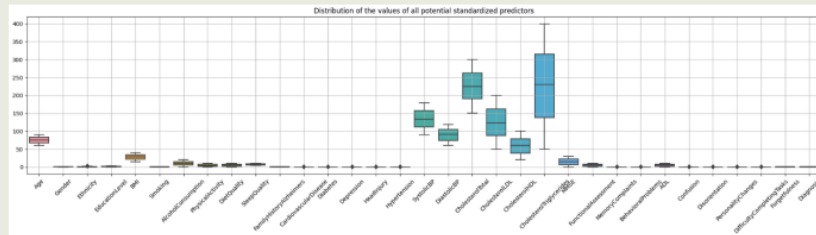
КОРЕЛЯЦІЙНА МАТРИЦЯ НАБОРУ ДАНИХ



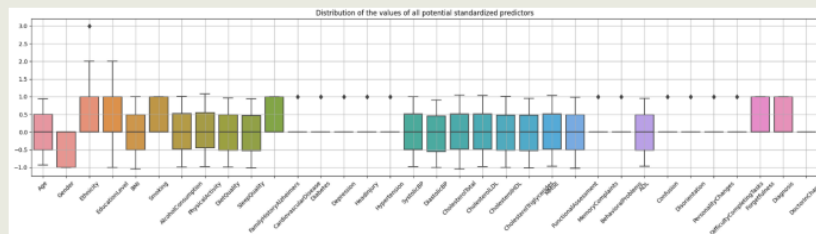
14

СТАНДАРТИЗАЦІЯ ДАНИХ

Розподіл значень усіх предикторів до стандартизації



Розподіл значень усіх предикторів після стандартизації за допомогою RobustScaler



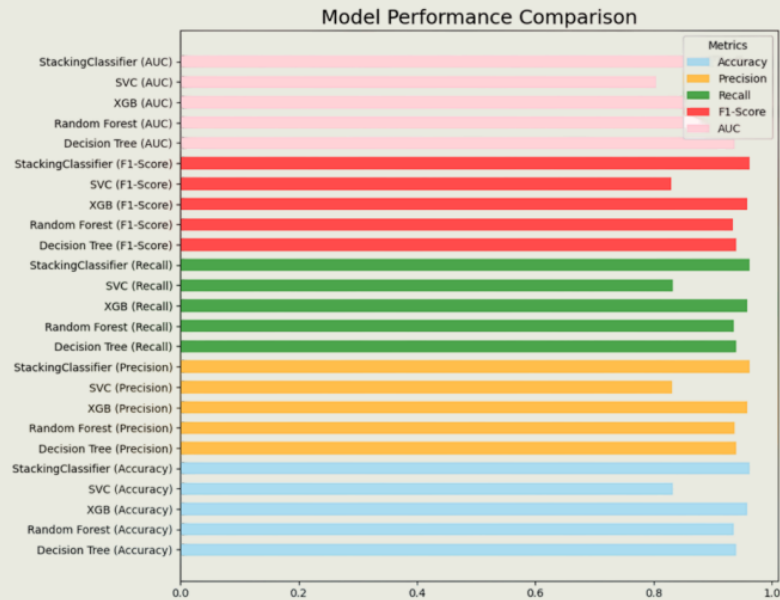
15

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

	Model	Accuracy	Precision	Recall	F1-Score	AUC
0	Decision Tree	0.939535	0.940049	0.939535	0.939705	0.936976
1	Random Forest	0.934884	0.936597	0.934884	0.933873	0.915811
2	XGB	0.958140	0.958140	0.958140	0.958140	0.954343
3	SVC	0.832558	0.830572	0.832558	0.829960	0.804205
4	StackingClassifier	0.962791	0.962791	0.962791	0.962791	0.959416

16

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ



17

ВИСНОВКИ

У результаті порівняльного аналізу класифікаційних моделей для прогнозування ризику хвороби Альцгеймера найкращі результати показав StackingClassifier, продемонструвавши найвищі значення Accuracy, Precision, Recall, F1-Score, AUC та найменшу кількість хибнопозитивних і хибнонегативних передбачень.

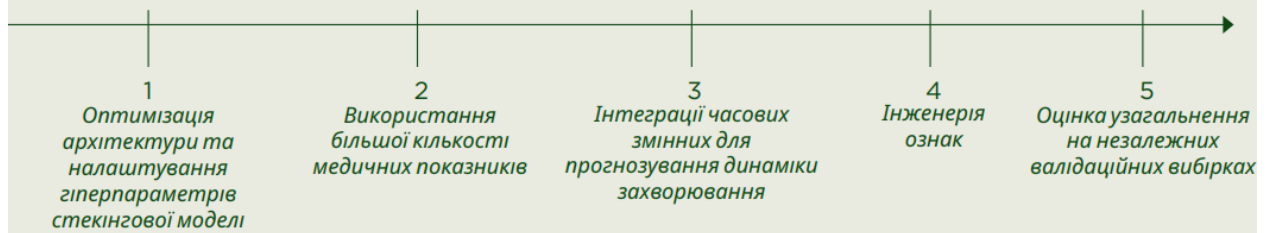
Accuracy	Precision	Recall	F1-Score	AUC
0.962791	0.962791	0.962791	0.962791	0.959416



Отже, StackingClassifier є оптимальним вибором завдяки своїй здатності узагальнювати закономірності в даних і збалансовано розпізнавати обидва класи, що особливо важливо у медичних задачах.

18

ПОДАЛЬШІ ДОСЛІДЖЕННЯ



Дякую за увагу!