

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ  
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ  
імені ІГОРЯ СІКОРСЬКОГО»  
ФАКУЛЬТЕТ БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ  
(повна назва інституту/факультету)

кафедра БІОМЕДИЧНОЇ КІБЕРНЕТИКИ  
(повна назва кафедри)

«На правах рукопису»  
УДК \_\_\_\_\_

«До захисту допущено»  
В.о.завідувача кафедри БМК

\_\_\_\_\_ Світлана АЛХІМОВА  
(підпис) (ініціали, ПРІЗВИЩЕ)

“ \_\_\_\_\_ ” *грудня* 2025р.

**Магістерська дисертація**  
на здобуття ступеня магістра  
за освітньо-професійною програмою **«Комп'ютерні технології в біології та медицині»**  
зі спеціальності 122 **«Комп'ютерні науки»**

на тему: **Система прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників**

Виконав: студент II курсу, групи ЗК-41мп

**СЕРЕДЕНКО АНДРІЙ ОЛЕКСАНДРОВИЧ**

(прізвище, ім'я, по батькові)



(підпис)

Науковий керівник: *доцент каф. біомедичної кібернетики (БМК)*  
*к.т.н., доцент, Федорін Ілля Валерійович*

(посада, науковий ступінь, вчене звання, прізвище, ім'я, по ініціали)



(підпис)

Рецензент: *професор каф. біомедичної інженерії (БМІ)*

*д.м.н., професор, Максименко Віталій Борисович*

(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали)

(підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.

Студент \_\_\_\_\_

(підпис)



Київ – 2025 року

**Національний технічний університет України**  
**«Київський політехнічний інститут імені Ігоря Сікорського»**  
**Факультет біомедичної інженерії**  
**Кафедра біомедичної кібернетики**

Рівень вищої освіти – другий (магістерський)

Спеціальність – 122 «Комп'ютерні науки»

Освітньо-професійна програма «Комп'ютерні технології в біології та медицині»

ЗАТВЕРДЖУЮ

В.о.завідувача кафедри БМК

\_\_\_\_\_ Світлана АЛХІМОВА

«\_\_\_\_» листопада 2025 р.

**ЗАВДАННЯ**  
**на магістерську дисертацію студента**

**СЕРЕДЕНКО АНДРІЙ ОЛЕКСАНДРОВИЧ**

(прізвище, ім'я, по батькові)

1. Тема дисертації \_\_\_\_\_ **Система прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників**

науковий керівник дисертації

***Федорін Ілля Валерійович, к.ф.-м.н., доцент каф. БМК***

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «03» листопада 2025 р. №4751-с\_

2. Термін подання студентом дисертації \_\_\_\_\_ **20-22 листопада 2025 року**

3. Об'єкт дослідження: *Проблемне використання Інтернету дітьми.*

4. Вихідні дані: *База даних Healthy Brain Network з даними демографічних, фізичних, психологічних та поведінкових даних учасників, включаючи часові ряди з акселерометрів та показники інтернет-залежності.*

5. Перелік завдань, які потрібно розробити: *Провести огляд існуючих рішень за темою МД. Визначити засоби реалізації алгоритму прогнозування. Обґрунтувати вибір технологічного стеку, провести тестування різних алгоритмів машинного навчання та обрати фінальну модель. Створити та протестувати систему прогнозування. Провести узагальнення результатів. Виконати розробку стартап-проєкту за темою МД. Зробити загальні висновки.*

6. Орієнтовний перелік графічного (ілюстративного) матеріалу: *14 слайдів (презентація з захисту МД).*

7. Орієнтовний перелік публікацій: *1 стаття у фаховому журналі категорії Б.*

8. Консультанти розділів дисертації<sup>1\*</sup>: *Не передбачено.*

9. Дата видачі завдання **28 серпня 2025р.**

<sup>1\*</sup> Якщо визначені консультанти. Консультантом не може бути зазначено наукового керівника магістерської дисертації.

## Календарний план

| № з/п | Назва етапів виконання магістерської дисертації                                                                                                         | Термін виконання етапів МД | Примітка        |
|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|-----------------|
| 1     | Отримати завдання на МД                                                                                                                                 | До 28.08.2025              | <i>виконано</i> |
| 2     | Завершення виконання практичної частина МД                                                                                                              | До 25.10.2025              | <i>виконано</i> |
| 3     | Завершення оформлення розділів МД (Вступ, Основні розділи та висновки до них, Загальні висновки, Список використаних джерел)                            | До 24.11.2025              | <i>виконано</i> |
| 4     | Апробація та публікація результатів дослідження МД (отримати підтвердження про прийняття до друку)                                                      | До 01.12.2025              | <i>виконано</i> |
| 5     | Перевірка МД науковим керівником                                                                                                                        | 24.11.2025                 | <i>виконано</i> |
| 6     | Подання в електронному вигляді МД на перевірку нормоконтролера                                                                                          | До 01.12.2025              |                 |
| 7     | Подання в електронному вигляді МД на перевірку подібності Strike Plagiarism com. Отримати позитивний звіт подібності.                                   | До 05.12.2025              |                 |
| 8     | Підготовка Експертної оцінки до звіту подібності.                                                                                                       | До 10.12.2025              |                 |
| 9     | Отримати відгук наукового керівника                                                                                                                     | 10.12.2025                 |                 |
| 10    | Надати на кафедру пакет документів в паперовому та електронному вигляді (МД, відгук керівника, звіт подібності, Експертної оцінки до звіту подібності ) | 10-11.12.2025              |                 |
| 11    | Отримати допуск до захисту МД в ЕК та направлення до рецензента (засідання кафедри)                                                                     | Засідання кафедри          |                 |
| 12    | Подання МД рецензенту. Отримати на надання рецензію до ЕК / на кафедру                                                                                  | До 19.12.2025              |                 |
| 13    | Подання супровідного пакету документів по МД до захисту в ЕК <sup>2</sup>                                                                               | До 19.12.2025              |                 |
| 14    | Захист МД в ЕК                                                                                                                                          | 22 – 26.12.2025            |                 |

Студент



(підпис)

**Андрій СЕРЕДЕНКО**

(ім'я, ПРІЗВИЩЕ)

Науковий керівник



(підпис)

**Ілля ФЕДОРІН**

(ім'я, ПРІЗВИЩЕ)

Нормоконтролер

(підпис)

**Галина КОРНІЄНКО**

(ім'я, ПРІЗВИЩЕ)

<sup>2</sup> не пізніше ніж за один тиждень до затвердженої дати захисту МД в ЕК

## РЕФЕРАТ

Магістерська дисертація за темою «Система прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників» виконана студентом кафедри біомедичної кібернетики ФБМІ *Середенко Андрієм Олександровичем* зі спеціальності 122 «Комп'ютерні науки» за освітньо-професійною програмою «Комп'ютерні технології в біології та медицині» та складається зі: вступу; 5 розділів (огляд матеріалів та існуючих рішень з прогнозування проблемного використання інтернету дітьми; засоби реалізації алгоритму прогнозування рівня проблемного використання інтернету дітьми; обґрунтування вибору технологічного та модельного стека для створення системи прогнозування рівня проблемного використання інтернету дітьми; програмна реалізація та методика роботи системи прогнозування рівня проблемного використання інтернету дітьми; узагальнення результатів роботи зі створення системи прогнозування рівня проблемного використання інтернету дітьми), розділу зі стартап проєкту, висновків до кожного з цих розділів; загальних висновків; списку використаних джерел, який налічує 41 джерел. Загальний обсяг роботи 138 сторінка.

**Актуальність теми.** Стрімкий ріст кількості випадків проблемного використання Інтернету серед молодого покоління, що веде до погіршення як психологічного та емоційного, так і фізичного стану, можуть нести негативні наслідки для населення всього світу. Наразі, класичні підходи до виявлення подібних станів, в своїй більшості спираються на анкети та опитувальники, та носять частково суб'єктивний характер, не враховуючи фізичні ознаки, які можуть стати важливими індикаторами. Поєднання такого роду даних з іншими, за використанням методів машинного навчання, відкриває широкі перспективи до створення інноваційної та унікальної системи прогнозування рівня проблемного використання Інтернету дітьми.

**Мета і задачі дослідження.**

Метою роботи є покращення наявних та виявлення нових підходів та методів до прогнозування рівня проблемного використання Інтернету дітьми,

використовуючи фізичні показники, за допомогою методів машинного навчання.

Сформульовано наступні **задачі**:

1. Аналіз наявних матеріалів та існуючих рішень за темою дослідження;
2. Огляд засобів реалізації прогнозувального алгоритму;
3. Обґрунтування вибору технологічного стека для практичної частини роботи;
4. Робота з набором даних;
5. Тестування алгоритмів машинного навчання, вибір фінальної моделі;
6. Програмна реалізація системи прогнозування;
7. Узагальнення результатів роботи;
8. Розробка стартап-проєкту за темою дослідження.

**Об'єкт дослідження.** Проблемне використання Інтернету дітьми.

**Предмет дослідження.** Прогнозування рівня проблемного використання Інтернету дітьми.

**Методи дослідження.** В процесі проведеного аналізу було визначено та обґрунтовано використання наступного інструментарія: мови програмування Python, її бібліотек (Pandas, NumPy, Scikit-learn, PyTorch, Skorch, Joblib, Matplotlib, Seaborn), фреймворку для створення графічного інтерфейсу користувача Tkinter, середовища розробки PyCharm, різноманітних методів, алгоритмів і підходів машинного навчання.

**Практичне значення одержаних результатів.** Магістерська дисертація виконана за ініціативної наукової роботи кафедри НДР д/р № 0123U100866 «Методи та моделі ідентифікації станів об'єктів в задачах прийняття медичних рішень» (06.02.2023). Робота може мати практичне застосування у сферах медицини, освіти, психології. Отриманий в результаті роботи програмний застосунок може бути інтегровано в медичні, освітні та виховні системи та платформи, а також в персоналізовані носимі пристрої для забезпечення можливості ранньої ідентифікації проблемного використання Інтернету дітьми і використовуватися лікарями, педагогами, батьками тощо.

**Апробація результатів дослідження.** Не заплановано.

**Публікації.** За результатами виконаної роботи була опублікована 1 наукова стаття в фаховому журналі категорії Б:

Середенко А.О., Федорін І.В. 2025. Моделювання ризику інтернет-залежності у дітей на основі мультипараметричних даних та методів машинного навчання. *Наука і техніка сьогодні*. 2025. № 11(52). с. 2835-2848. DOI: [https://doi.org/10.52058/2786-6025-2025-11\(52\)](https://doi.org/10.52058/2786-6025-2025-11(52)) (дата звернення 15.12.2025).

**Ключові слова.** Проблемне використання Інтернету, Інтернет-залежність, діти та підлітки, фізичні показники, машинне навчання, система прогнозування.

***Бібліографічний опис МД***

Середенко, А.О. Система прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників: магістерська дис.: 122 Комп'ютерні науки / Середенко Андрій Олександрович. – Київ, 2025. – 138 с

## ABSTRACT

The master's dissertation on the topic «System for predicting the level of problematic Internet use by children based on physical indicators» was completed by *Andrii Seredenko, a student of the Department of Biomedical Cybernetics of the FBME. The work was carried out within the framework of the educational-professional program «Computer Technologies in Biology and Medicine» under the specialization of 122 «Computer Science».* The dissertation consists of: introduction; 6 chapters (*review of materials and existing solutions for predicting problematic internet use in children; tools for implementing the algorithm for predicting the level of problematic internet use in children; justification of the choice of the technological and model stack for building a system for predicting the level of problematic internet use in children; software implementation and operating methodology of the system for predicting the level of problematic internet use in children; summary of the results of developing the system for predicting the level of problematic internet use in children; startup project based on the topic of the master's research*), a section on the startup project, conclusions for each of these chapters; general conclusions; list of sources used, which includes 41 sources and appendices. Total volume of the work is 138 pages.

**Actuality of theme.** The rapid growth of problematic Internet use among the younger generation, which leads to a deterioration in both psychological and emotional and physical health, can have negative consequences for the population of the whole world. Currently, classical approaches to identifying such conditions are mostly based on questionnaires and questionnaires, and are partly subjective in nature, not taking into account physical signs that can become important indicators. Combining this type of data with others, using machine learning methods, opens up broad prospects for creating an innovative and unique system for predicting the level of problematic Internet use by children.

**The purpose and tasks of the work.**

*The purpose* of the work is to improve existing and identify new approaches and methods for predicting the level of problematic Internet use by children, using physical indicators, using machine learning methods.

The following **tasks** are formulated:

1. Analysis of available materials and existing solutions on the topic of the study;
2. Overview of means of implementing the predictive algorithm;
3. Justification of the choice of a technological stack for the practical part of the work;
4. Working with a data set;
5. Testing machine learning algorithms, choosing the final model;
6. Software implementation of the forecasting system;
7. Generalization of the results of the work;
8. Development of a startup project on the topic of the study.

**Object of research.** Problematic use of the Internet by children.

**Subject of research.** Predicting the level of problematic use of the Internet by children.

**Used methods.** During the analysis, the use of the following tools was identified and justified: the Python programming language, its libraries (Pandas, NumPy, Scikit-learn, PyTorch, Skorch, Joblib, Matplotlib, Seaborn), the Tkinter graphical user interface framework, the PyCharm development environment, and various machine learning methods, algorithms, and approaches.

**Practical significance of the obtained results.** The master's dissertation was carried out within the initiative research project of the department, R&D № 0123U100866 «Methods and models for identifying the states of objects in medical decision-making tasks» (06.02.2023). The work has potential practical applications in medicine, education, and psychology. The software application developed in the course of the thesis can be integrated into medical, educational, and pedagogical systems and platforms, as well as into personalized wearable devices, enabling early identification of problematic Internet use in children. It may be used by healthcare professionals, educators, parents, and other stakeholders.

**Approbation of the research results.** Not planned.

***Publications.*** Based on the results of the work, 1 scientific article was published in a category B journal:

Seredenko A., Fedorin I. 2025. Modeling the risk of Internet addiction in children based on multiparametric data and machine learning methods. *Science and Technology Today*. 2025 No. 11(52), pp. 2835-2848. DOI: [https://doi.org/10.52058/2786-6025-2025-11\(52\)](https://doi.org/10.52058/2786-6025-2025-11(52)) (access date 15.12.2025).

***Keywords.*** Problematic Internet use, Internet addiction, children and adolescents, physical indicators, machine learning, forecasting system.

## ЗМІСТ

|                                                                                                                                                                                        |           |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ.....</b>                                                                                                         | <b>14</b> |
| <b>ВСТУП.....</b>                                                                                                                                                                      | <b>15</b> |
| <b>РОЗДІЛ 1 ОГЛЯД МАТЕРІАЛІВ ТА ІСНУЮЧИХ РІШЕНЬ З ПРОГНОЗУВАННЯ ПРОБЛЕМНОГО ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ.....</b>                                                                     | <b>19</b> |
| 1.1. Огляд теоретичних основ за тематикою дослідження.....                                                                                                                             | 19        |
| 1.1.1. <i>Проблемне використання Інтернету дітьми .....</i>                                                                                                                            | <i>19</i> |
| 1.1.2. <i>Підходи до діагностики проблемного використання Інтернету .....</i>                                                                                                          | <i>21</i> |
| 1.1.3. <i>Використання фізичних показників у задачі прогнозування проблемного використання Інтернету .....</i>                                                                         | <i>22</i> |
| 1.1.4. <i>Технології прогнозування проблемного використання Інтернету .....</i>                                                                                                        | <i>24</i> |
| 1.2. <i>Наявні дослідження та підходи до прогнозування проблемного використання Інтернету дітьми .....</i>                                                                             | <i>26</i> |
| 1.2.1. <i>Дослідження «Оцінка психічного стресу серед студентів коледжу з використанням даних про частоту серцевих скорочень та прискорення рук, зібраних з носимих датчиків».....</i> | <i>26</i> |
| 1.2.2. <i>Дослідження «Використання машинного навчання для дослідження факторів ризику проблемного використання смартфонів серед підлітків».....</i>                                   | <i>27</i> |
| 1.2.3. <i>Дослідження «Застосування машинного навчання для прогнозування поведінкової залежності підлітків від Інтернету» .....</i>                                                    | <i>28</i> |
| 1.2.4. <i>Дослідження «Прогнозування проблемного використання Інтернету дітьми».....</i>                                                                                               | <i>29</i> |
| Висновки до розділу 1 .....                                                                                                                                                            | 30        |
| <b>РОЗДІЛ 2 ЗАСОБИ РЕАЛІЗАЦІЇ АЛГОРИТМУ ПРОГНОЗУВАННЯ РІВНЯ ПРОБЛЕМНОГО ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ.....</b>                                                                         | <b>32</b> |
| 2.1. Базові алгоритми.....                                                                                                                                                             | 32        |

|                                                                                                        |    |
|--------------------------------------------------------------------------------------------------------|----|
| 2.1.1. Лінійна регресія (Linear Regression) .....                                                      | 32 |
| 2.1.2. Логістична регресія (Logistic Regression) .....                                                 | 33 |
| 2.1.3. Древа рішень (Decision Tree) .....                                                              | 35 |
| 2.2. Просунуті поодинокі моделі .....                                                                  | 36 |
| 2.2.1. Метод <i>K</i> -найближчих сусідів (K-Nearest Neighbors) .....                                  | 36 |
| 2.2.2. Машина опорних векторів (Support Vector Machine) .....                                          | 38 |
| 2.3. Ансамблеві підходи .....                                                                          | 39 |
| 2.3.1. Ансамблеве навчання та його типи .....                                                          | 39 |
| 2.3.2. Оптимізація ансамблю .....                                                                      | 42 |
| 2.3.3. Випадковий ліс (Random Forest) .....                                                            | 42 |
| 2.3.4. Градієнтний бустинг (Gradient Boosting) та його реалізації<br>XGBoost, LightGBM, CatBoost ..... | 44 |
| 2.4. Нейромережеві моделі .....                                                                        | 46 |
| 2.5. Схема експериментів і валідації .....                                                             | 48 |
| 2.5.1. Розбиття даних .....                                                                            | 48 |
| 2.5.2. Метрика якості коефіцієнту згоди Каппа .....                                                    | 49 |
| Висновки до розділу 2 .....                                                                            | 50 |

## **РОЗДІЛ 3 ОБГРУНТУВАННЯ ВИБОРУ ТЕХНОЛОГІЧНОГО ТА МОДЕЛЬНОГО СТЕКА ДЛЯ СТВОРЕННЯ СИСТЕМИ ПРОГНОЗУВАННЯ РІВНЯ ПРОБЛЕМНОГО ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ..... 51**

|                                                           |    |
|-----------------------------------------------------------|----|
| 3.1. Обґрунтування вибору технологічного стека .....      | 51 |
| 3.2.1. Вибір мови програмування .....                     | 52 |
| 3.2.2. Вибір бібліотек .....                              | 53 |
| 3.2.3. Вибір середовища розробки .....                    | 55 |
| 3.2. Обґрунтування вибору та робота з набором даних ..... | 55 |
| 3.2.1. Загальний огляд датасету .....                     | 56 |
| 3.2.2. Робота з даними актиграфії .....                   | 57 |
| 3.2.3. Робота з табличними даними .....                   | 64 |
| 3.2.4. Оброблення пропущених значень .....                | 78 |
| 3.2.5. Видалення надлишкових ознак .....                  | 80 |
| 3.2.6. Визначення важливості ознак .....                  | 84 |

|                                                               |            |
|---------------------------------------------------------------|------------|
| 3.3. Обґрунтування вибору фінальної моделі .....              | 84         |
| 3.3.1. Загальна схема роботи .....                            | 85         |
| 3.3.2. Класифікаційні моделі .....                            | 86         |
| 3.3.3. Регресійні моделі.....                                 | 89         |
| 3.3.4. Нейронні мережі.....                                   | 92         |
| 3.3.5. Фінальна модель.....                                   | 95         |
| Висновки до розділу 3 .....                                   | 97         |
| <b>РОЗДІЛ 4 ПРОГРАМНА РЕАЛІЗАЦІЯ ТА МЕТОДИКА РОБОТИ</b>       |            |
| <b>СИСТЕМИ ПРОГНОЗУВАННЯ РІВНЯ ПРОБЛЕМНОГО</b>                |            |
| <b>ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ .....</b>                    | <b>98</b>  |
| 4.1. Функціональні вимоги до програмного застосунку.....      | 98         |
| 4.2. Програмна реалізація програмного застосунку.....         | 99         |
| 4.2.1. Архітектура програмного застосунку.....                | 99         |
| 4.2.2. Потік даних під час роботи програмного застосунку..... | 100        |
| 4.2.3. Тестування роботи програмного застосунку .....         | 101        |
| 4.3. Методика роботи програмного застосунку.....              | 101        |
| Висновки до розділу 4 .....                                   | 106        |
| <b>РОЗДІЛ 5 УЗАГАЛЬНЕННЯ РЕЗУЛЬТАТІВ РОБОТИ ЗІ</b>            |            |
| <b>СТВОРЕННЯ СИСТЕМИ ПРОГНОЗУВАННЯ РІВНЯ ПРОБЛЕМНОГО</b>      |            |
| <b>ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ .....</b>                    | <b>107</b> |
| 5.1. Аналіз отриманих результатів .....                       | 107        |
| 5.2. Унікальність проведеного дослідження .....               | 109        |
| 5.2.1. Схема роботи з даними.....                             | 109        |
| 5.2.2. Фінальна конфігурація моделі .....                     | 110        |
| 5.2.3. Порівняння результатів роботи моделі з наявними        |            |
| дослідженнями .....                                           | 111        |
| 5.3. Практичне значення результатів .....                     | 111        |
| Висновки до розділу 5 .....                                   | 112        |
| <b>РОЗДІЛ 6 СТАРТАП-ПРОЄКТ ЗА ТЕМОЮ МАГІСТЕРСЬКОГО</b>        |            |
| <b>ДОСЛІДЖЕННЯ .....</b>                                      | <b>114</b> |
| 6.1. Короткий опис проєкту .....                              | 114        |

|                                            |            |
|--------------------------------------------|------------|
| 6.2. Бізнес-модель .....                   | 115        |
| 6.2.1. Цінність та вимоги до продукту..... | 115        |
| 6.2.2. Сегмент споживачів .....            | 117        |
| 6.2.3. Канали збуту.....                   | 118        |
| 6.2.4. Взаємодія зі споживачами.....       | 119        |
| 6.2.5. Дохід (монетизація) .....           | 120        |
| 6.2.6. Ключові види діяльності .....       | 121        |
| 6.2.7. Ключові ресурси .....               | 122        |
| 6.2.8. Ключові партнери.....               | 123        |
| 6.2.9. Витрати .....                       | 123        |
| 6.2.10. Споживчі властивості товару.....   | 124        |
| 6.2.11. Дослідження ринку .....            | 125        |
| 6.2.12. Елементи фінансового плану.....    | 126        |
| 6.2.13. Резюме.....                        | 127        |
| Висновки до розділу 6 .....                | 128        |
| <b>ЗАГАЛЬНІ ВИСНОВКИ .....</b>             | <b>129</b> |
| <b>СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ .....</b>    | <b>132</b> |

## **ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ**

МД – магістерська дисертація

ПП – програмний продукт

ПЗ – програмний застосунок

ЧСС – частота серцевих скорочень

## ВСТУП

*Актуальність* поставленої задачі МД, що має на увазі створення системи прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників, в цей момент часу є питанням високої значущості.

У поточних реаліях світу цифрових технологій, мабуть, найбільшим з «китів», на яких він тримається, є мережа Інтернет. Легкість та швидкість навчання, комунікацій, різноматність способів проведення часу – беззаперечно величезні переваги сучасного життя, які можливі лише завдяки мережі. Водночас із цим, Інтернет може призводити і до негативних наслідків для психологічного та навіть фізичного здоров'я людей. Це підтверджує і зростаюча в геометричній прогресії кількість наукових досліджень, що однозначно стверджують: надмірне, безконтрольне використання мережі у великій кількості випадків може призвести до появи залежності або ж так званого «проблемного використання Інтернету» (Problematic Internet Use) [1].

Ріст кількості негативних наслідків проблемного використання Інтернету відповідно призводить до зростання необхідності створення спеціалізованих методів та підходів до виявлення та профілактики даного явища. Традиційними методами вважаються психологічні тести та опитувальники, що лише в незначній мірі вирішують проблему. Такий підхід є суб'єктивним, оскільки не враховує велику кількість різноманітних факторів, серед яких – дослідження фізіологічних змін в тілі людини [2].

Проблемне використання Інтернету наразі можна пов'язати з наступними факторами: зниженням фізичної активності протягом дня, розладами зі сном, змінами в серцево-судинній системі тощо, що і підтверджується численними дослідженнями. Таким чином, використання всіх цих чинників в якості індикаторів проблемного використання мережі може відкривати нові, обширні горизонти [3].

Найбільшою групою ризику наразі є молоде покоління, до якого належать діти та підлітки. З кожним роком мінімальний вік, в якому діти

починають користуватися цифровими технологіями та Інтернетом, знижується. У ранньому віці діти проявляють більшу схильність до зовнішніх впливів, ніж сформовані особистості, що і є основним фактором формування залежної поведінки [4].

Методи машинного навчання – один з найбільш популярних підходів до реалізації задач прогнозування та передбачення, в тому числі й виявленні індикаторів певних явищ, зокрема проблемного використання Інтернету. Вони містять в собі можливість обробки великих масивів різноманітних (фізіологічних, психологічних, поведінкових тощо) даних, виявляючи приховані, нелінійні закономірності та будуючи на основі цього якісні та точні моделі. Поєднання алгоритмів машинного навчання з фізичними даними (наприклад, з носимих пристроїв, що неперервно збирають інформацію), відкриває перспективи до створення гібридних систем раннього виявлення проблеми, що в подальшому зможе покращити психологічне та фізичне здоров'я зростаючого покоління [2; 5].

Таким чином, ця робота спрямована на дослідження перспектив використання сучасних методів машинного навчання до розв'язання задачі прогнозування рівня проблемного використання мережі Інтернет дітьми, із використанням фізичних показників.

### ***Мета і завдання роботи.***

*Метою* роботи є удосконалення наявних та виявлення нових підходів та методів прогнозування рівня проблемного використання Інтернету дітьми, на основі фізичних показників, за допомогою методів машинного навчання.

Її досягнення передбачає вирішення наступних *завдань*:

1. Аналіз наявних матеріалів та існуючих рішень за темою дослідження;
2. Огляд засобів реалізації прогнозувального алгоритму;
3. Обґрунтування вибору технологічного стека для практичної частини роботи;
4. Робота з набором даних;
5. Тестування алгоритмів машинного навчання, вибір фінальної моделі;

6. Програмна реалізація системи прогнозування;
7. Узагальнення результатів роботи;
8. Розробка стартап-проєкту за темою дослідження.

**Використані методи.** В процесі проведеного аналізу було визначено та обґрунтовано використання наступного інструментарія: мови програмування Python, її бібліотек (Pandas, NumPy, Scikit-learn, PyTorch, Skorch, Joblib, Matplotlib, Seaborn), фреймворку для створення графічного інтерфейсу користувача Tkinter, середовища розробки PyCharm, різноманітних методів, алгоритмів і підходів машинного навчання.

**Отримані результати.** Робота може мати практичне застосування у сферах медицини, освіти, психології. Отриманий в результаті роботи ПЗ може бути інтегровано в медичні, освітні та виховні системи та платформи, а також в персоналізовані носимі пристрої для забезпечення можливості ранньої ідентифікації проблемного використання Інтернету дітьми і використовуватися лікарями, педагогами, батьками тощо.

**Апробація результатів роботи.** Не заплановано.

**Публікації.** За результатами виконаної роботи було подано 1 наукову статтю до публікації у фаховому журналі категорії Б:

**Публікації.** За результатами виконаної роботи була опублікована 1 наукова стаття в фаховому журналі категорії Б:

Середенко А.О., Федорін І.В. 2025. Моделювання ризику інтернет-залежності у дітей на основі мультипараметричних даних та методів машинного навчання. *Наука і техніка сьогодні*. 2025. № 11(52). с. 2835-2848. DOI: [https://doi.org/10.52058/2786-6025-2025-11\(52\)](https://doi.org/10.52058/2786-6025-2025-11(52)) (дата звернення 15.12.2025).

**Структура роботи.**

Магістерська дисертація за темою «Система прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників» виконана студентом *Середенко Андрієм Олександровичем* зі спеціальності

122 «Комп'ютерні науки» за освітньо-професійною програмою «Комп'ютерні технології в біології та медицині», побудована за класичним типом та викладена на 138 сторінці машинописного тексту. Вона складається з: вступу; 6 розділів (огляд матеріалів та існуючих рішень з прогнозування проблемного використання інтернету дітьми; засоби реалізації алгоритму прогнозування рівня проблемного використання інтернету дітьми; обґрунтування вибору технологічного та модельного стека для створення системи прогнозування рівня проблемного використання інтернету дітьми; програмна реалізація та методика роботи системи прогнозування рівня проблемного використання інтернету дітьми; узагальнення результатів роботи зі створення системи прогнозування рівня проблемного використання інтернету дітьми; стартап-проект за темою магістерського дослідження), висновків до кожного з цих розділів; загальних висновків; списку використаних джерел, який налічує 41 джерело (1 – на кирилиці, 40 – на латиниці). В роботі представлено 83 рисунки.

## **РОЗДІЛ 1**

### **ОГЛЯД МАТЕРІАЛІВ ТА ІСНУЮЧИХ РІШЕНЬ З ПРОГНОЗУВАННЯ ПРОБЛЕМНОГО ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ**

Цей розділ охоплює огляд літературних джерел різних видів, що знаходяться у відкритому доступі (починаючи зі звичайних онлайн-ресурсів і до наукових досліджень та статей) за тематикою прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників, а також проводиться аналіз вже наявних підходів та рішень.

#### **1.1. Огляд теоретичних основ за тематикою дослідження**

Найпершим, критично необхідним кроком, є огляд теоретичних основ за тематикою дослідження для глибшого розуміння контексту поставленої задачі, можливих напрямів для подальшого практичного виконання роботи.

##### **1.1.1. Проблемне використання Інтернету дітьми**

Від початку двадцять першого сторіччя дедалі сильніше спостерігається зростання ролі мережі Інтернет у житті більшої частини населення нашої планети, та, зокрема, молодого покоління. Широка доступність цифрових технологій відкриває все більші можливості для багатьох сфер їх життя. Разом із тим як науковці, так і медичні організації все частіше підіймають питання ризиків, які виникають внаслідок надмірного та безконтрольного використання мережі Інтернет, що потенційно становить загрозу фізичному та психоемоційному здоров'ю дітей [1; 3; 4].

Наукові джерела називають вищеописане явище такими термінами: «Problematic Internet Use (PIU)», «Internet Addiction Disorder», «Excessive Internet Use». Також для окремих випадків (пов'язаних з ігровою залежністю), застосовується термін «Gaming Disorder». Важливим уточненням є те, що Всесвітня організація охорони здоров'я у своїй Міжнародній класифікації

хвороб (одинадцятого перегляду) офіційно додала ігровий розлад у діагностичні категорії, тим самим підкреслюючи його актуальність у межах сучасної медицини та психології [6; 7].

Залежність від Інтернету визначається як психічний розлад, який характеризується нав'язливим прагненням безперервно користуватися мережею в будь-яких її проявах, а також неможливістю припинення процесу онлайн-активності [1; 6]. Таким чином відбувається зміщення уваги з реального життя на віртуальне. У такому стані, людина здатна повністю ігнорувати всі свої базові потреби, як соціальні, так і фізіологічні, тим самим створюючи досить серйозні ризики для свого життя та благополуччя [1; 3].

Звужуючи увагу досліджень саме до сегменту дітей та підлітків, науковці дійшли висновку, що саме ця вікова категорія населення перебуває у найбільшій зоні ризику, у контексті формування залежності від мережі Інтернет. До основних чинників, що спричиняють подібні розлади відносять незрілість особистості, схильність до суттєвого впливу новизни, соціальний тиск середовища, а також нерозвиненість критичного мислення [3; 4].

Епідеміологічні дослідження доводять актуальність обраної тематики, показуючи суттєві значення поширеності проблемного використання мережі Інтернет серед дітей – від 6% до 18%, залежно від країни і регіону, методів оцінки та варіативності вибірки. Для європейських країн, середнім є значення, близьке до 12%, в той час коли країни Азії досить часто досягають значень, більших за 20%. Це явище зазвичай виникає через різний рівень цифровізації країн, а також специфікою культури використання сучасних технологій та, зокрема, мережі Інтернет [8].

Якщо говорити про Україну, з часом проблема теж актуалізується. Таким чином, за дослідженням Інституту соціальної та політичної психології НАПН України, вже більше 15% дітей та підлітків, віком від дванадцяти до сімнадцяти років, проявляють ознаки проблемного використання Інтернету. Найбільшу загрозу, наразі, несуть соціальні мережі, комп'ютерні ігри та платформи для відображення відеоматеріалів [9].

Підсумовуючи, проблемне використання Інтернету в молоді можна визначити як різносторонній феномен, який охоплює психологічні, поведінкові, фізіологічні та соціальні фактори. В наш час, поширеність даної тематики, ґрунтується на обширній кількості досліджень, що вказують на кардинально негативний вплив неконтрольованого використання мережі на здоров'я та благополуччя дітей, вже зараз руйнуючи їх майбутнє. Актуальність нашої роботи також спирається і на відсутність об'єктивно та коректно працюючих аналогів системи діагностики та передбачення вищевказаного явища [1; 3; 4; 5].

### **1.1.2. Підходи до діагностики проблемного використання Інтернету**

Основним підходом до виявлення проблемного використання мережі Інтернет, наразі є використання різноманітних психометричних тестів та опитувальників, за якими досліджується рівень розвиненості проблеми, та які заповнюються самими учасниками. Нижче наведено найпоширеніші з них [1; 10; 11; 12]:

1. Internet Addiction Test (IAT) – анкета, що складається з двадцяти пунктів, спрямованих на аналіз неконтрольованості використання мережі, проблем соціалізації та міжособистісних стосунків, рівня емоційної залученості до реального життя тощо. Цей опитувальник є найбільш використовуваним у великій кількості країн світу, включаючи як клінічні, так і наукові та дослідницькі організації [1].

2. Chen Internet Addiction Scale (CIAS) – анкета, що складається з двадцяти шести пунктів, спрямованих на аналіз проблем соціалізації та міжособистісних стосунків, рівня залежності від мережі, проблем з тайм-менеджментом тощо. Найчастіше цей опитувальник використовується в країнах Азії [10].

3. Generalized Problematic Internet Use Scale 2 (GPIUS2) – анкета, що складається з п'ятнадцяти пунктів, спрямованих на аналіз когнітивних та соціально-емоційних проблем, що виникають внаслідок проблемного

використання мережі Інтернет. Цей опитувальник також застосовується у багатьох країнах по всьому світу [11].

4. Problematic Internet Use Questionnaire (PIUQ) – анкета, що складається з вісімнадцяти пунктів на оцінку наявності та рівня залежності від користування Інтернетом, та аналізом за трьома основними категоріями: надмірне та проблемне використання мережі, наслідки для соціального життя та повноцінного функціонування [12].

Ключовою перевагою використання даних інструментів є можливість стандартизації результатів, що дозволяє пришвидшувати отримання результатів та виявлення ризиків, порівнюючи отримані дані між різноманітними організаціями та країнами, проводячи спільні та масштабні наукові дослідження [1; 10; 11; 12].

Попри вищенаведені позитивні аспекти, існують і суттєві недоліки. Найбільший – це суб'єктивність власної оцінки учасника експерименту, що може значно занижувати якість отримуваних результатів. Мовні та культурні особливості, рівень інтегрованості інструментів також можуть впливати на інтерпретацію та точність оцінок [1; 10; 11; 12]. Разом із цим, такі анкети не спрямовані на врахування взаємозв'язку фізичних показників із проблемним використанням Інтернету, що суттєво звужує можливості раннього виявлення проблем [3; 5].

Саме тому, фахівці різних галузей все частіше звертають увагу на потребу поєднання фізичних даних з іншими у процесі виявлення проблемного використання мережі Інтернет, тим самим підвищуючи якість та точність прогнозування, зменшуючи вплив суб'єктивних факторів на результати [2; 5; 13].

### **1.1.3. Використання фізичних показників у задачі прогнозування проблемного використання Інтернету**

Вищезазначена аргументація та опис сучасних підходів до виявлення проблемного використання Інтернету, що спираються у своїй більшості на

суб'єктивні опитувальники провокує ріст інтересу до використання об'єктивніших показників, зокрема фізичних. Такі дані можуть містити приховані зв'язки та слугувати гарними індикаторами для виявлення надмірного та неконтрольованого використання мережі, її впливу на організм. Використання такого підходу до прогнозування посилить надійність та швидкість отримання результатів для своєчасного розпізнавання станів підвищеного ризику [2; 3; 5; 13; 14; 15; 16; 17].

Відповідно, важливим може бути врахування наступних чинників:

1. Відомості про сон та циркадні ритми. Завищене використання мережі Інтернет, особливо ввечері, часто призводить до проблем зі сном (неспокійний сон, зменшення тривалості, неможливість засинання, постійна втома та відсутність відчуття відновлення). Наукові дослідження доводять, що у дітей проблемне використання часто супроводжується так званим «джетлагом», порушенням циркадних ритмів, що негативно впливає на загальний психоемоційний стан та когнітивні можливості організму [3; 4].

2. Рівень фізичної активності. Підвищений час використання цифрових технологій логічно знижує рівень фізичної активності, та підвищує кількість часу в сидячому положенні, що підтверджується рядом досліджень. Діти та підлітки, в яких було діагностовано високий рівень залежності, зазвичай мають набагато нижчу рухову активність протягом дня, тим самим погіршуючи своє фізичне здоров'я [3; 14]. Це і сприяє росту популярності використання носимих пристроїв – смарт-годинників, фітнес-браслетів, які стають невід'ємним «помічником» життєдіяльності людини, оскільки створюють можливість відстеження та контролю таких показників фізичної активності, як кількість кроків, рівень спожитих ресурсів, метаболічні показники, що може стати надійним індикатором змін у поведінці [2; 13; 16]. Ожиріння, що є наслідком малорухливого способу життя, також можна певною мірою пов'язати з надлишковим використанням мережі [3].

3. Серцево-судинні показники та вегетативна регуляція. Інтернет-залежність, окрім вищезазначених чинників, може супроводжуватися

зростанням симпатичної активності нервової системи та порушеннями варіабельності ЧСС. Дослідження, що проводяться серед молоді, вказують на тісний зв'язок зниження варіативності серцевого ритму та проблемним використанням Інтернету, що підтверджується наявністю стресового стану та розладами вегетативної регуляції організму. Крім того ЧСС в нормальному, спокійному стані зростає, діагностуючи перевантаження фізіологічних систем [3; 14].

Окрім цього всього, перевищення індивідуальної норми часу використання Інтернету досить часто призводить до серйозних проблем із зором (таких як короткозорість та сухість очей), внаслідок високого навантаження на очі [3].

Враховуючи всю описану інформацію, використання фізичних даних при виявленні та прогнозуванні проблемного використання Інтернету значно розширює перспективи створення відповідних моделей передбачення, що безпосередньо може вплинути на більш раннє та точне діагностування проблемних індикаторів, забезпечуючи максимальну ефективність напрямків моніторингу та профілактики [2; 5; 13; 16; 17].

#### **1.1.4. Технології прогнозування проблемного використання Інтернету**

Зважаючи на багатогранність природи виникнення розладів, пов'язаних із проблемним використанням Інтернету, сучасні дослідження все більше звертаються до комп'ютерних методів обробки даних, моделей машинного навчання та нейромереж, які здатні працювати з великими масивами інформації та виявляти глибоко приховані взаємозв'язки між ознаками, прогножуючи досить точні та якісні значення. Наразі найбільш актуальним серед дослідників є напрям поєднання психологічних анкет та опитувальників з фізичними, фізіологічними та статистичними ознаками, тим самим формуючи повноцінний та комплексний шлях до виявлення проблемного використання Інтернету [5; 14; 15].

Таким чином, найбільш популярними методами є:

1. Методи та алгоритми машинного навчання та нейронні мережі. Використовуються як для класифікації одразу на класи, так для регресії, що прогнозують певний числовий показник, який у подальшому переводиться у фінальний результат за завчасно встановленими межами. Такий підхід має багато переваг, серед яких є можливість інтеграції з носимими пристроями, для точної фіксації певних фізичних показників [2; 5; 14; 16].

2. Інтеграція носимих технологій. Як було зазначено вище, у сучасному світі дедалі частіше використовуються такі пристрої, як фітнес-браслети, смарт-годинники тощо, із вбудованими спеціальними сенсорами, що використовуються для збору даних у реальному часі. Інформація щодо сну, циркадних ритмів, фізичної активності, серцевого ритму та його частоти – все це, в комбінації з результатами психологічних тестів, можуть стати ключовими індикаторами Інтернет-залежності та іншими пов'язаними з нею розладами [2; 13; 16].

3. Великі масиви даних «Big Data» та цифрові сліди. Наукові дослідження вказують на пряму кореляцію тривалості, частоти, контенту, що переглядається та його різновидів (соцмережі, ігри та інші сервіси) з виявленням розладів. Саме тому використання алгоритмів та методів «Big Data» розширює можливості ідентифікації груп ризику (проте важливо пам'ятати про питання конфіденційності даних) [17; 18].

Комбінуючи всі ці пункти, формуються адаптивні та інтегровні інструменти, що поєднують у собі аналіз великих масивів психологічних, фізіологічних та поведінкових даних. Вони можуть приймати вигляд мобільних додатків та освітніх сервісів, надаючи батькам та педагогам надійний та точний висновок щодо наявності індикаторів проблемного використання Інтернету у їхніх дітей та учнів [2; 5; 14; 16; 17].

## **1.2. Наявні дослідження та підходи до прогнозування проблемного використання Інтернету дітьми**

Для більш повного розуміння рівня розвитку обраного напрямку дослідження, є необхідним аналіз і оцінка вже існуючих підходів до прогнозування проблемного використання Інтернету дітьми. Саме тому в цьому пункті проводиться огляд літератури та зазначених там алгоритмів вирішення поставленої задачі.

### **1.2.1. Дослідження «Оцінка психічного стресу серед студентів коледжу з використанням даних про частоту серцевих скорочень та прискорення рук, зібраних з носимих датчиків»**

Яскравим прикладом використання методів машинного навчання у поєднанні з набором фізичних даних є дослідження, проведене на студентах коледжу в США. Його метою стало використання носимих пристроїв у задачі оцінки рівня психічного стресу, з інтеграцією в мобільний додаток [13].

Учасниками експерименту стали п'ятдесят чотири студенти коледжу, у віці від вісімнадцяти до тридцяти чотирьох років, що протягом 40-денного періоду не знімали з себе розумні годинники, із вмонтованими в них спеціальними сенсорами. Пристрій автоматично зчитував дані ЧСС та прискорення рухів руки. У той самий час, кожен студент мав змогу власноруч вносити інформацію щодо набутого стресу, використовуючи спеціалізований мобільний додаток Mhelp [13].

За час експерименту було отримано понад три тисячі стресових ситуацій, і близько тридцяти тисяч контрольних спостережень. У дослідженні була поставлена задача класифікації, із використанням таких методів машинного навчання, як випадковий ліс, машина опорних векторів, просунутий бустинговий метод XGBoost та інші. Саме XGBoost показав найвищу точність із значенням 84.5% та площею під кривою AUC (0.64) [13].

Ознаками, які найбільше вплинули на результати точності передбачень стали показники мінімального значення ЧСС, варіабельності серцевого ритму та стандартне відхилення середнього значення прискорення руки [13].

У результаті дослідження, було отримано висновки, про доцільність використання фізіологічних даних щодо серцево-судинної системи та параметрів моторики частин тіла, в контексті прогнозування емоційного стану та рівня стресу людини, зокрема молодого віку [13].

Таким чином, це дослідження, на основі своїх результатів, дає надію на подальший розвиток не тільки у контексті задачі прогнозування психічного стресу, а й багатьох інших, у тому числі і виявлення проблемного використання Інтернету та Інтернет-залежності у дітей, базуючись на тому, що вони супроводжуються як психоемоційними, так і фізіологічними змінами в організмі. Тому, можна зі впевненістю стверджувати, що дана робота слугує гарним прикладом впровадження сучасних технологій для передбачення психологічних та поведінкових розладів [13].

#### **1.2.2. Дослідження «Використання машинного навчання для дослідження факторів ризику проблемного використання смартфонів серед підлітків»**

Наступним, варто розглянути дослідження, засноване на аналізі проблемного використання смартфонів Problematic Smartphone Use (PSU) серед групи канадських підлітків, яке набуло актуальності під час пандемії COVID-19. Воно є ближчим за тематикою до обраної, і так само ґрунтується на використанні методів машинного навчання [14].

Учасниками експерименту стали дві тисячі сто учнів зі шкіл Британської Колумбії у віці від одинадцяти до вісімнадцяти років. Кожен із них мав скласти онлайн-опитувальник, заснований на короткій версії шкали Smartphone Addiction Scale (SAS). Окрім цього проводився аналіз рівнів саморегуляції (психо-емоційної, поведінкової, когнітивної) та страху втрати цифрових пристроїв, екранного часу тощо [14].

На відміну від попереднього дослідження, у цьому перед науковцями було поставлено регресійну задачу із використанням таких алгоритмів як випадковий ліс, просунутий бустинговий метод XGBoost та різноманітних shrinkage-методів, які й показали найближчі до реальних значень результатів, а найкращим став алгоритм Lasso, з найбільш збалансованими результатами метрик якості RMSE (0.656), MAE (0.519) та  $R^2$  (0.34) на тесті [14].

Найважливішими ознаками для прогнозування стали: страх втрати доступу до девайсів та мережі (FoMO), та показники саморегуляції різних рівнів. Натомість, найбільш логічний, на перший погляд, екранний час, проведений із гаджетами не дав суттєвого покращення для результуючих моделей, що й підтверджує важливість виявленн прихованих взаємозв'язків під час передбачення [14].

Підсумовуючи, це дослідження вказує на потребу використання варіативності підходів до дослідження питань раннього виявлення індикаторів проблемного використання мережі Інтернет, та гаджетів загалом, а також застосування машинного навчання у межах цієї тематики [14].

### **1.2.3. Дослідження «Застосування машинного навчання для прогнозування поведінкової залежності підлітків від Інтернету»**

Наступним є дослідження, вже напряму пов'язане з обраною тематикою та поставленою задачею, а саме вивчення залежності від мережі Інтернет серед підлітків за допомогою машинного навчання [5].

Учасниками експерименту стали чотири тисячі чотириста шістьдесят один учень старших курсів з міста Чунцін. За тестом Інтернет-залежності, високий рівень було виявлено у тисячі двохсот десяти підлітків. До датасету увійшли різні види інформації – від демографічних та соціальних даних, до різноманітних психологічних опитувальників для оцінки стану учасника [5].

У дослідженні тестувалася робота таких методів машинного навчання, як багатосаровий перцептрон, випадковий ліс, K-найближчих сусідів, машина опорних векторів, логістична регресія та просунутий бустинговий метод XGBoost, серед яких найвищу точність показав алгоритм логістичної регресії,

зі значенням площі під кривою AUC (0.847). Конкурентний результат продемонстрував метод XGBoost, а метод найближчих сусідів став найгіршим серед представлених [5].

Ознаками, що можна вважати безпосередніми маркерами залежності від інтернету, виявились стать, країна походження учасника, ознаки імпульсивної та невротичної поведінки, психічних розладів та депресивного стану, що супроводжувалися проблемами з ворожістю до оточуючих [5].

У результаті аналізу отриманих значень метрик оцінки якості роботи, машинне навчання вкотре стало надійним інструментом, показавши достатньо точні прогнози поведінкової залежності від мережі Інтернет, поєднуючи всі наявні інформаційні джерела. У той самий час демонструється необхідність періодичного моніторингу психологічного стану молодих людей за використанням сучасних технологій [5].

#### **1.2.4. Дослідження «Прогнозування проблемного використання Інтернету дітьми»**

Прямим аналогом та конкурентом нашого дослідження є роботи, проведені у рамках змагання «Child Mind Institute – Problematic Internet Use», за тематикою якого й була поставлена задача [15; 19].

Учасниками експерименту стали близько трьох тисяч молодих людей з Нью-Йорку, у віці від п'яти до двадцяти двох років. Усі дані були поєднані в спільний датасет під егідою одного з найбільших досліджень, пов'язаних із психологічним здоров'ям молоді Healthy Brain Network, що включив у себе як фізіологічну інформацію (антропометрія, сон, фізична активність, актиграфічні показники), так і результати традиційних анкетних методів [15; 19].

Змінною для прогнозування став індекс Severity Impairment Index (SII), що напряду визначає рівень проблемного використання Інтернету серед досліджуваних. Цей індекс містив чотири рівні вираженості залежності, від «відсутності» симптоматики до «сильно вираженої» [15; 19].

Велика кількість пропущених значень (у тому числі і серед ключових змінних), серйозний дисбаланс класів, а також складність роботи з певними видами даних – основні виклики, з якими довелося зіткнутися. У роботі, окрім звичайних методів машинного навчання та нейромереж, було розглянуто такі підходи, як алгоритми відновлення (заповнення) пропущеної інформації, нормалізацію даних, додавання нових змінних, та інші підходи [15].

Найбільше значення основної метрики оцінки якості роботи в дослідженні (коефіцієнта Каппи) на тесті показала нейронна мережа (0.354), хоча це є слабким результатом узгодженості отриманих прогнозів із реальними даними [15].

Озаками виявилися поведінкові та антропометричні характеристики та змінні, пов'язані зі сном [15].

У результаті, було створено підґрунтя для подальшої роботи, розкривши деякі нюанси дослідження, а також довівши актуальність нашої роботи з прогнозування рівня проблемного використання мережі Інтернет дітьми [15].

## **Висновки до розділу 1**

У світі, де цифрові технології набувають дедалі більшого значення, питання залежності від гаджетів та мережі Інтернет, а також проблемності їх використання, зокрема серед дітей та підлітків, набувають всесвітнього масштабу. Проблемне використання Інтернету складається з численної кількості факторів, поєднуючи психологічні та поведінкові, соціальні та фізіологічні аспекти життя.

Традиційні методи діагностики, в основі яких лежать звичайні анкети, тестування та опитування певною мірою є зручними та стандартизованими інструментами, проте занадто суб'єктивними, як для сфер психології та медицини.

З іншого ж боку, фізіологічні показники (відомості про сон, ЧСС та варіабельність серцевого ритму, рівень фізичної активності) вже зараз

доводять свою цінність як надійні маркери виявлення проблемного використання Інтернету.

Сучасні роботи, що наразі доступні у відкритому доступі, здебільшого використовують методи та алгоритми машинного навчання, поєднуючи їх із даними носимих пристроїв, тим самим інтегруючи збір різноманітних даних у фінальні спеціалізовані моделі, що значно підвищує якість і точність при ранній діагностиці, виявленню та профілактиці проблем з цифровими технологіями у молодого покоління.

## **РОЗДІЛ 2**

# **ЗАСОБИ РЕАЛІЗАЦІЇ АЛГОРИТМУ ПРОГНОЗУВАННЯ РІВНЯ ПРОБЛЕМНОГО ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ**

Цей розділ містить розгорнутий аналіз та опис різних методів машинного навчання: прості базові алгоритми, просунуті поодинокі моделі, ансамблеві підходи та нейромережі, їхні покрокові стратегії, сильні та слабкі сторони. Наведено схему поділу даних та описано метрику оцінки якості роботи алгоритмів.

### **2.1. Базові алгоритми**

Почнемо з розгляду найпростіших базових алгоритмів машинного навчання. Розглянуто три алгоритми – лінійна та логістична регресії, дерева рішень.

#### **2.1.1. Лінійна регресія (Linear Regression)**

Лінійна регресія – один з базових і найпростіших, але водночас із тим найбільш використовуваних алгоритмів у машинному навчанні з вчителем, що використовується у контексті регресійних задач для прогнозування числових значень. За її допомогою можна легко виявити лінійні зв'язки між ознаками, тим самим отримуючи фінальне рішення в простіших задачах, або ж будуючи підґрунтя для подальшої роботи складніших [20].

Загальна схема роботи алгоритму лінійної регресії:

1. Визначення числової ключової ознаки (ціль передбачення) та набору змінних, що можуть мати на неї вплив (за якими виконується передбачення);
2. Створення початкової неоптимізованої моделі;
3. Тестування роботи моделі, оцінка якості та точності її роботи;
4. Підбір параметрів для покращення роботи моделі, її оптимізація;

5. Тестування фінальної версії моделі, оцінка якості та точності її роботи на нових даних, що не брали участі в навчанні.

Візуальне відображення схеми роботи алгоритму (рис. 2.1):

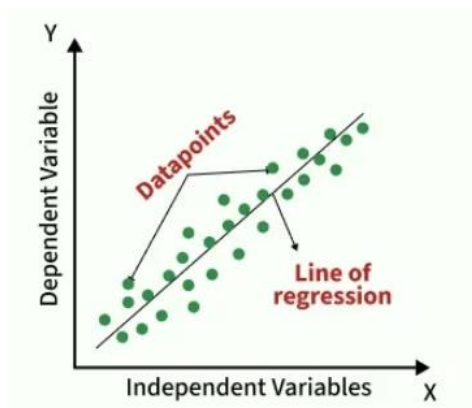


Рисунок 2.1 – Алгоритм роботи Linear Regression [20]

Основними перевагами використання алгоритму лінійної регресії є простота застосування та інтерпретованість результатів, потреба малої кількості ресурсів, а також можливість отримати початкове уявлення про зв'язки між ознаками [20].

Недоліками ж є неможливість гарантування точності результатів при роботі з даними, що мають глибокі та приховані, зокрема нелінійні зв'язки, чутливість до викидів, пропусків значень, неоднорідності даних та їх ненормалізованості [20].

Підсумовуючи, для роботи з прогнозування рівня проблемного використання Інтернету дітьми, лінійна регресія може бути чудовим початковим інструментом отримання базового розуміння та виявлення лінійних залежностей змінних, що може стати у нагоді для подальшого аналізу.

### 2.1.2. Логістична регресія (Logistic Regression)

Логістична регресія – ще один базовий алгоритм машинного навчання з учителем, що також є досить популярним та застосовується для вирішення задач з класифікації. Основною його задачею є розподіл екземплярів з даними по класах (бінарна – два класи, мультикласова – декілька класів) [21].

Загальна схема роботи алгоритму логістичної регресії:

1. Визначення категоріальної ключової ознаки (ціль передбачення) та набору змінних, що можуть мати на неї вплив (за якими виконується передбачення);
2. Створення початкової, неоптимізованої моделі;
3. Тестування роботи моделі, оцінка якості та точності її роботи;
4. Підбір параметрів для покращення роботи моделі, її оптимізація;
5. Тестування фінальної версії моделі, оцінка якості та точності її роботи на нових даних, що не брали участі у навчанні.

Візуальне відображення схеми роботи алгоритму (рис. 2.2):

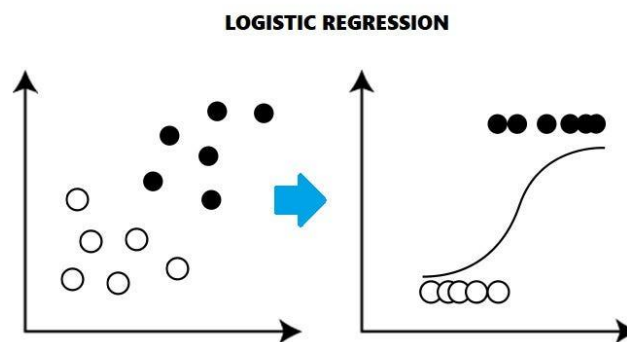


Рисунок 2.2 – Алгоритм роботи Logistic Regression [21]

Як і у випадку лінійної, алгоритм роботи логістичної регресії характеризується простотою та інтерпретованістю, низькою обчислювальною вартістю на великих об'ємах даних, а також малим ризиком перенавчання [21].

З недоліків можна виділити – обмеженість до застосування лише до задач класифікації, припущення щодо лінійності зв'язків між змінними, а також висока чутливість до викидів, пропусків значень, неоднорідності даних та їх ненормалізованості [21].

При роботі з поставленою задачею прогнозування психологічних розладів внаслідок проблемного використання мережі Інтернет, логістична регресія може стати першим кроком у створенні повноцінної класифікаційної моделі.

### 2.1.3. Древа рішень (Decision Tree)

Древа рішень – останній з базових алгоритмів машинного навчання з учителем, що будуть розглядатися. Цей підхід може застосовуватися як для задач класифікації, так і для задач регресії. Головною ідеєю є покрокове розділення всіх наявних даних за завчасно визначеними умовами (що і створює при візуалізації форму дерева). Вузли становлять собою умови, гілки – варіанти рішень, листи – результати (для задачі класифікації – належність до певного класу, а для регресії – числове значення) [22].

Загальна схема роботи алгоритму дерева рішень (рис. 2.3):

1. Створення головного вузла зі всіма даними (корінь дерева);
2. Визначення початкової ознаки поділу за певним критерієм якості;
3. Покроковий поділ даних відповідно до умови обраної ознаки;
4. Ітеративне повторення процесу вибору ознаки та розділення для кожної з підмножин, поки не буде досягнуто умови для зупинки (глибина дерева, досягнення розділу на класи, відсутність покращення результатів тощо). Нероздільні вузли називають листками з фінальними рішеннями.

Візуальне відображення схеми роботи алгоритму (рис. 2.3):

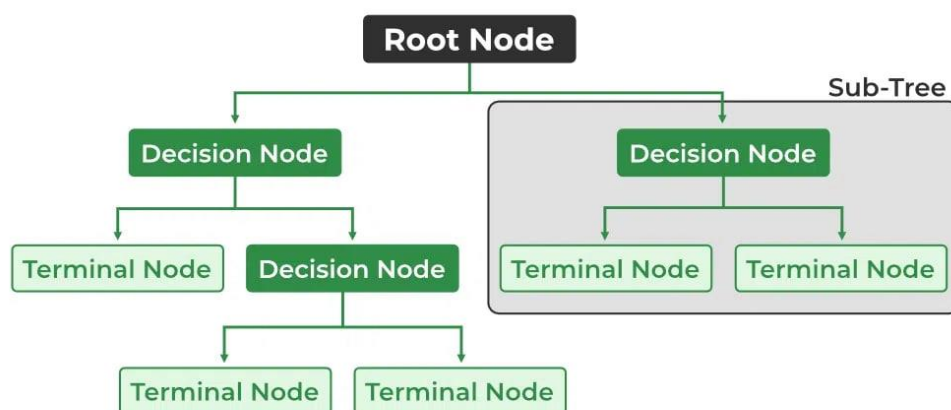


Рисунок 2.3 – Алгоритм роботи Decision Tree [22]

Перевагами використання дерев рішень, як і для всіх інших базових алгоритмів, є простота та інтерпретованість, можливість оцінки важливості

змінних, а також робота з числовими та класовими даними, в умовах відсутності препроцесингу [22].

З недоліків – висока ймовірність перенавчання, за відсутності належного критерію зупинки роботи алгоритмів, а також відсутність стабільності результатів внаслідок навіть незначних змін [22].

У результаті, для задачі прогнозування рівня проблемного використання Інтернету дітьми, такий підхід може допомогти в визначення важливості характеристик, внаслідок підпорядкуванню простим правилам, стаючи корисним інструментом для лікарів, батьків та вчителів.

## **2.2. Просунуті поодинокі моделі**

Наступним кроком є огляд більш просунутих поодиноких моделей машинного навчання, які вже самостійно можуть стати повноцінними інструментами для виявлення Інтернет-залежності та пов'язаних із нею проблем.

### **2.2.1. Метод К-найближчих сусідів (K-Nearest Neighbors)**

Метод К-найближчих сусідів – черговий метод машинного навчання з учителем, що в переважній більшості випадків застосовується в якості класифікатора, хоча й бувають регресійні винятки. Методика роботи алгоритму полягає у виявленні схожості, або ж «близькості» поточного екземпляра даних з навчальною вибіркою [23].

Загальна схема роботи алгоритму К-найближчих сусідів:

1. Визначення типу задачі та формування простору ознак;
2. Вибір кількості сусідів для роботи;
3. Вибір критерію близькості залежно від простору ознак;
4. Для кожного з екземплярів даних вираховується значення критерія близькості до інших екземплярів;
5. Визначення найближчих сусідів (з найменшими значеннями близькості);

6. Обчислення середнього (інколи зваженого) значення у задачі регресії, або ж визначення класу, спираючись на більшість найближчих сусідів до екземпляра.

Візуальне відображення схеми роботи алгоритму (рис. 2.4):

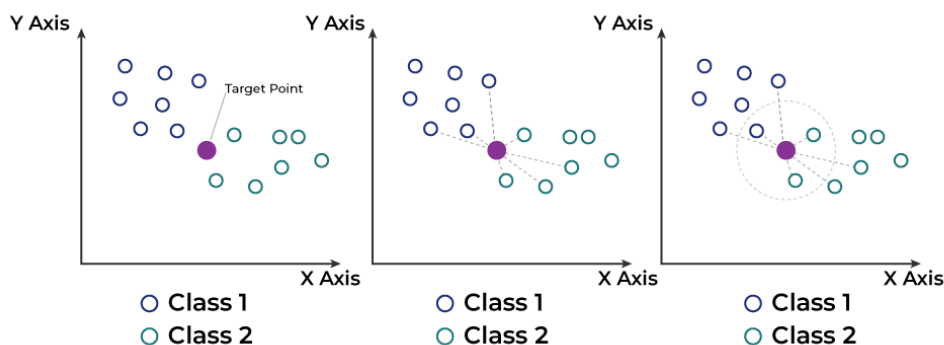


Рисунок 2.4 – Алгоритм роботи K-Nearest Neighbors [23]

Вищеописаний алгоритм, незважаючи на більшу складність, у порівнянні з базовими алгоритмами відзначається своєю простотою та зрозумілістю. Окрім цього, внаслідок відсутності процесу навчання має змогу працювати з різноманітними видами як числових, так і класових змінних, без урахування характеру взаємозв'язків між ознаками [23].

Поряд із цим, такий підхід вимушений зберігати більшу частину вибірки, що може досить сильно збільшувати час роботи алгоритму на великих обсягах інформації. Також, метод може бути чутливим до різного роду викидів, шумів та пропущених значень і потребує на вхід невеликої розмірності даних [23].

Таким чином, метод найближчих сусідів, зі своєю здатністю до розподілу за класами (а в даному випадку за групами ризику), може стати потужним інструментом в задачі передбачення рівня проблемного користування мережею Інтернет дітьми, як самостійно, так і в поєднанні з іншими підходами.

### 2.2.2. Машина опорних векторів (Support Vector Machine)

Метод машини опорних векторів – черговий метод машинного навчання з учителем, застосовний у задачах регресії, та, особливо, класифікації (з використанням гіперплощини для розмежування екземплярів за класами). Варто зазначити існування ядрових функцій – це допоміжні перетворення для випадків, коли неможливо лінійно визначити границю класів) [24].

Загальна схема роботи алгоритму машини опорних векторів (рис. 2.5):

1. Визначення типу задачі та формування простору ознак;
2. Вибір ядрової функції для роботи;
3. Створення гіперплощини, що поділяє дані на класи та виділення опорних векторів (екземпляри, найближчі за відстанню до площини) для класифікації та налаштування спеціалізованих параметрів, призначених для встановлення значення «жорсткості» припущень моделі щодо розташування даних відносно гіперплощини для регресії;
4. Оцінка якості та узагальненості роботи моделі.

Візуальне відображення схеми роботи алгоритму (рис. 2.5):

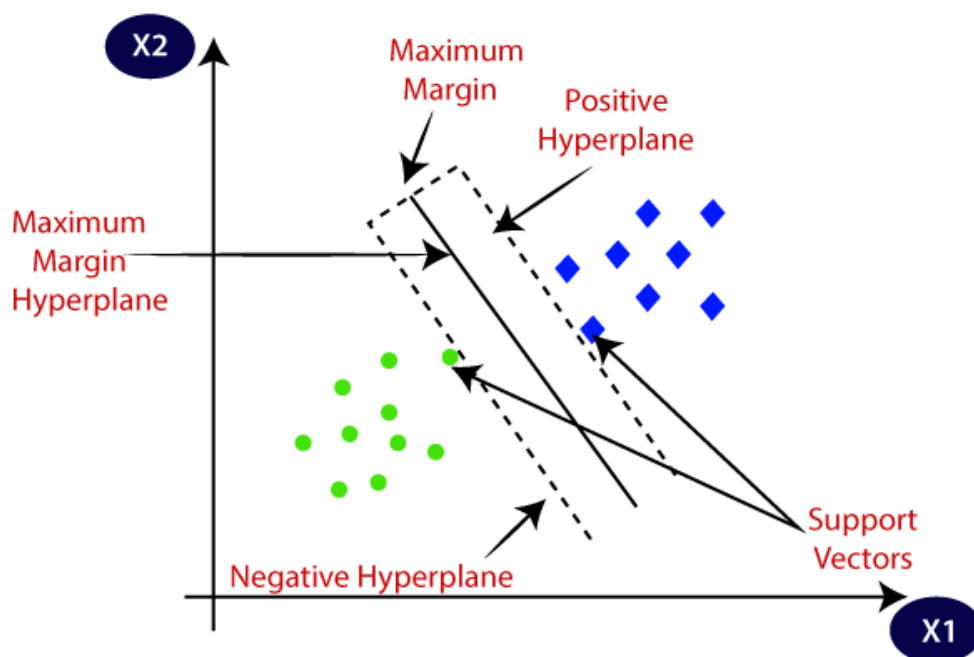


Рисунок 2.5 – Алгоритм роботи Support Vector Machine [24]

Перевагами описаного алгоритму є ефективність роботи з чітко визначеними рамками класів та значень, зокрема на широкій кількості ознак (чого не можна сказати про метод найближчих сусідів). Також метод може виявляти глибоко приховані нелінійні залежності в змінних, за допомогою різних типів ядер [24].

Недоліки – зі зростанням кількості інформації, зростає і обчислювальна складність, налаштування гіперпараметрів є трудомістким процесом, результати відчутно залежать від зашумленості даних та викидів [24].

У контексті нашої роботи, алгоритм машини опорних векторів є одним з обов'язкових для тестування підходів, завдяки своїй можливості до знаходження нелінійних взаємозв'язків, наприклад між психологічними тестовими даними та фізичними характеристиками, враховуючи багатокомпонентність явища проблемного використання Інтернету та Інтернет-залежності в дітей.

### **2.3. Ансамблеві підходи**

Унаслідок вивчення наукової літератури та досліджень було підтверджено точність прогнозів, отриманих на виході з ансамблевих алгоритмів машинного навчання (зокрема в межах дослідження Інтернет-залежності), тому однозначно доречним буде розгляд як самого поняття «ансамблю», його типів та можливих способів оптимізації, так і конкретних методів та підходів.

#### **2.3.1. Ансамблеве навчання та його типи**

Ансамблеве навчання – це клас методів машинного навчання, що об'єднує в собі кілька слабких базових алгоритмів у одну повноцінну модель. Таким чином досягається стабільність, знижується похибка результатів та зростає загальна якість роботи моделі, завдяки використанню декількох методів замість єдиного [25].

Ансамблевий підхід реалізується навчанням окремих моделей на унікальних підвибірках даних, або ж на різних вхідних параметрах, внаслідок чого, враховуючи результати роботи кожного з підходів на виході формується фінальне передбачення. Таким чином сильні сторони одних алгоритмів можуть перекривати слабкі інших, та навпаки [25].

Різні типи ансамблів працюють за своїми унікальними схемами:

1. Bagging – після навчання моделей на так званих «бутстреп-вибірках» із заміною, їх передбачення агрегуються (зазвичай в задачі регресії використовується середнє значення, в задачі класифікації – голосування) (рис. 2.6).

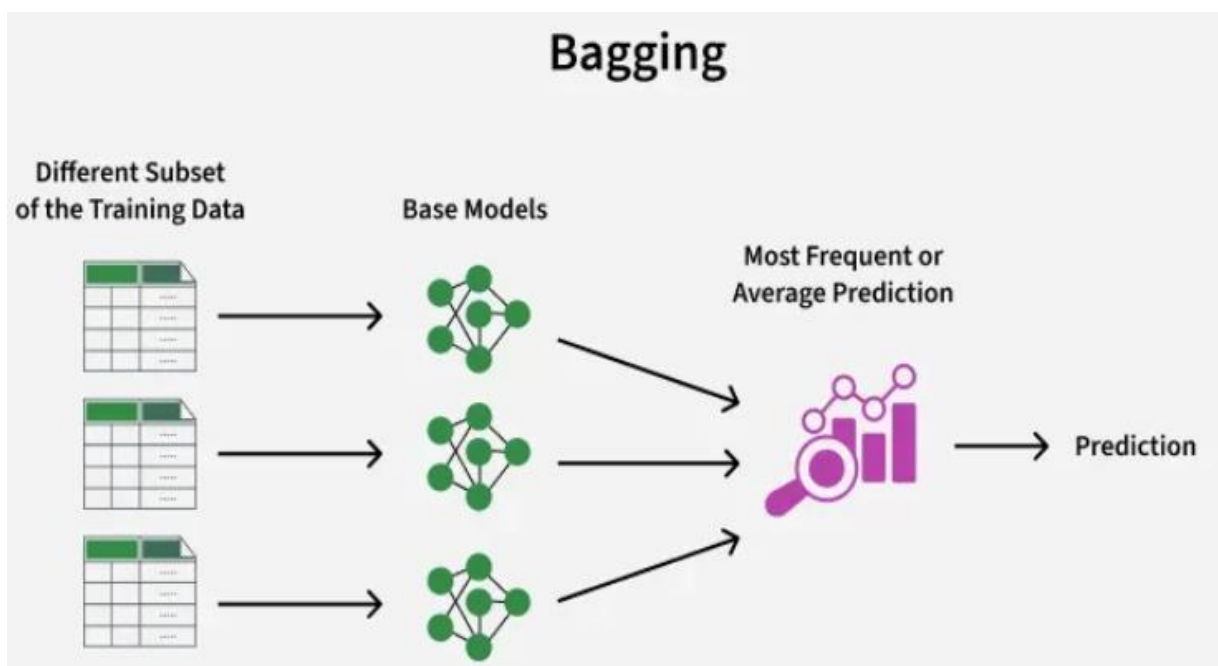


Рисунок 2.6 – Алгоритм роботи Bagging ансамблю [25]

2. Boosting – послідовне навчання моделей з урахуванням помилок попередніх, внаслідок чого коригується точність отримуваних результатів, що агрегуються з урахуванням важливості кожного з алгоритмів (більш сильні моделі у пріоритеті) (рис. 2.7).

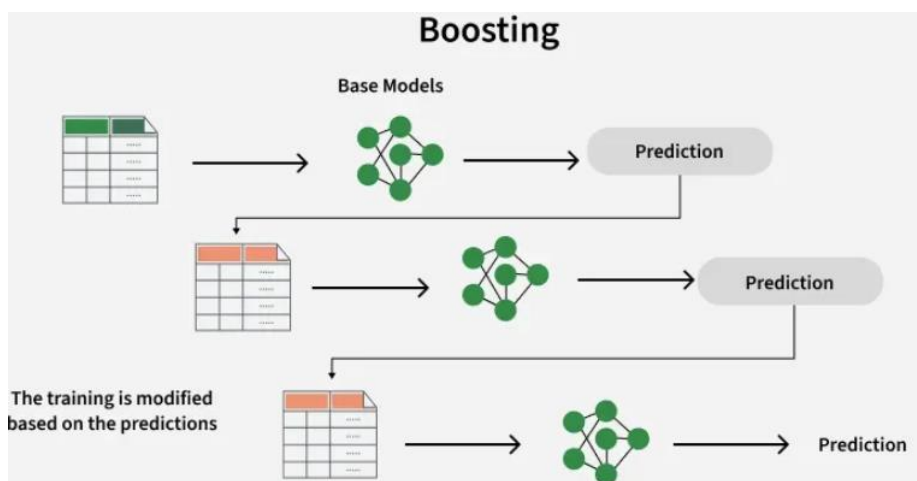


Рисунок 2.7 – Алгоритм роботи Boosting ансамблю [25]

3. Stacking – набір базових моделей різних типів проходить процес навчання на спільних даних та незалежно одна від одної, передбачення яких стають ознаками мета-моделі, що комбінує отримувані результати у фінальне значення (рис. 2.8).

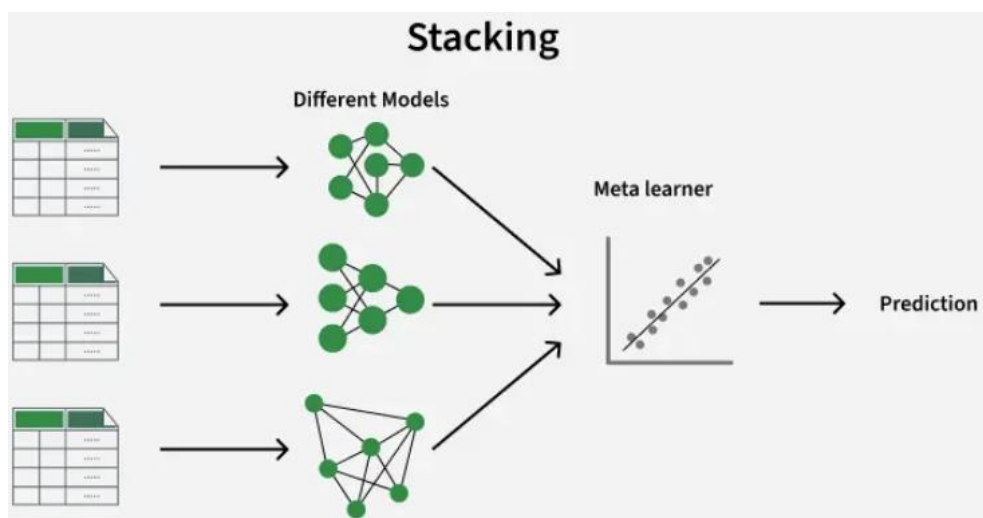


Рисунок 2.8 – Алгоритм роботи Stacking ансамблю [25]

Ансамблеві рішення є потужним інструментом, що забезпечує високу точність та якість отримуваних результатів, у порівнянні з поодинокими алгоритмами, збільшуючи шанс уникнення перенавчання, поєднуючи сильні сторони різноманітних підходів в єдину модель [25].

З іншої сторони вони потребують значної кількості обчислювальних ресурсів, складніші для розуміння та створення, а також важчі з точки зору інтерпретації результатів [25].

Повертаючись до поставленої задачі прогнозування, ансамблі мають значний потенціал для використання, знижуючи випадковість та вплив шумів і пропущених значень на фінальні прогнози, тим самим підвищуючи їх точність.

### **2.3.2. Оптимізація ансамблю**

Процес визначення найкращого типу ансамблю, складу алгоритмів і їх кількість, формування правил комбінування, підбору ваг та гіперпараметрів, методів регуляризації та ранньої зупинки, з метою покращення точності результатів – усе це називають оптимізацією ансамблю. Основною ідеєю вищеописаного процесу є визначення ідеального балансу між варіативністю та упередженням, забезпечуючи гнучкість та точність.

Основними кроками оптимізації ансамблю здебільшого є:

1. Вибір типу ансамблю та методу формування фінальних передбачень;
2. Формування набору базових моделей та визначення їх кількості;
3. Підбір гіперпараметрів кожної з моделей;
4. Оптимізація важливості моделей у результуючому ансамблі;
5. Застосування регуляризації та критеріїв ранньої зупинки роботи ансамблю;
6. Оцінка ресурсних витрат, точності та якості роботи ансамблю за різних умов.

### **2.3.3. Випадковий ліс (Random Forest)**

Метод випадкового лісу – перший конкретний ансамблевий метод машинного навчання, який буде розглянуто. Належить до ансамблів Bagging типу та застосовується як для задач класифікації, так і для задач регресії. В його основі лежить поєднання певної кількості дерев рішень (вже описаних у

роботі), процес навчання яких для більшої стабільності проходить незалежно – на випадковій підвибірці даних [26].

Загальна схема роботи алгоритму випадкового лісу:

1. Поділення датасету на підвибірки із заміною випадковим чином;
2. Побудова дерева рішень для кожної з підвбірок з випадково обраним піднабором ознак;
3. Проведення незалежного процесу навчання;
4. Агрегація результатів усередненням (регресія) або голосуванням (класифікація);
5. Валідація та оцінка якості роботи моделі.

Візуальне відображення схеми роботи алгоритму (рис. 2.9):

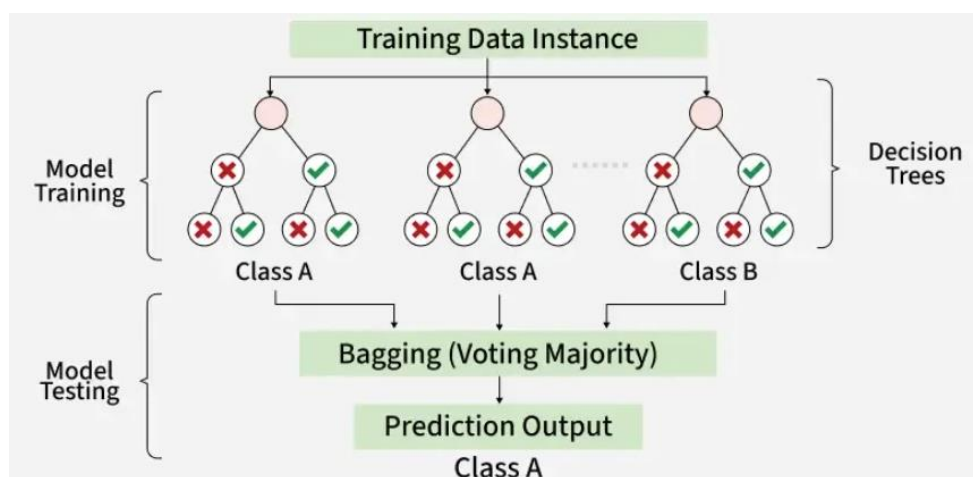


Рисунок 2.9 – Алгоритм роботи Random Forest [26]

Основними перевагами методу випадкового лісу є можливість роботи з різними типами задач прогнозування, гарна узагальненість інформації, що зменшує ризик виникнення перенавчання, стійкість до зашумленості даних та викидів [26].

Низька інтерпретованість прогнозів, висока обчислювальна вартість (внаслідок зростання кількості даних) та складність налаштування гіперпараметрів – основні недоліки методу [26].

Метод випадкового лісу завдяки своїм перевагам може стати цінним механізмом при побудові фінальної моделі для прогнозування індикаторів

проблемного використання Інтернету дітьми, забезпечуючи стабільність прогнозів за обширного простору ознак.

### 2.3.4. Градієнтний бустинг (Gradient Boosting) та його реалізації XGBoost, LightGBM, CatBoost

Принцип градієнтного бустингу є одним з найбільш популярних та потужних методів ансамблевого машинного навчання бустингового типу. Основою є покрокове додавання слабких моделей з корекцією помилок на кожному з кроків, з мінімізацією при цьому функції втрат через градієнтний спуск. Таким чином набувається здатність виявлення прихованих нелінійних взаємозв'язків [27].

Окрім базового методу градієнтного бустингу, основними реалізаціями принципу є XGBoost, LightGBM та CatBoost, що мають свої імплементації, налаштовані під різні конкретні види задач [28; 29; 30].

XGBoost – алгоритм, що базується на принципі градієнтного бустингу, підсумовуючи передбачення всіх дерев (рис. 2.10). Швидкість і точність отримуваних результатів на великих масивах даних – основні переваги методу [28].

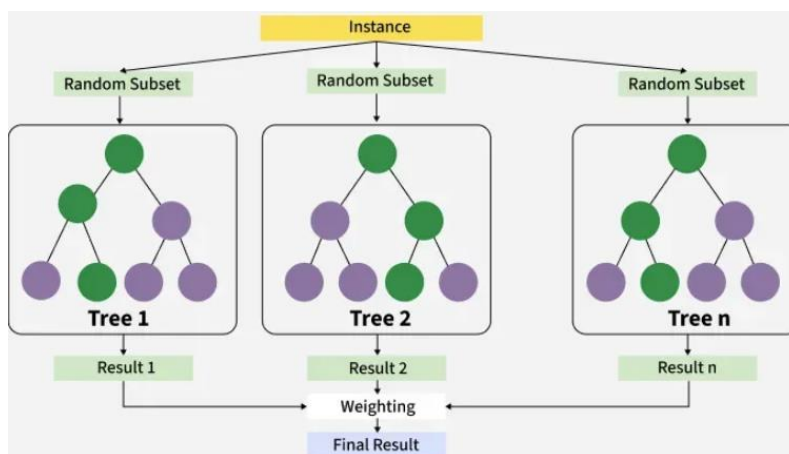


Рисунок 2.10 – Алгоритм роботи XGBoost [28]

LightGBM – алгоритм нового покоління за принципом градієнтного бустингу, що працює за стратегією «leaf-wise», де враховується раціональність позиції додавання нового вузла до листка (рис. 2.11). Хоча обчислювальні

витрати і можуть сягати суттєвих значень, існують додаткові оптимізації для їх зменшення, що паралельно підвищують ефективність роботи. Підтримує паралельність навчання та використання ресурсів графічного процесора [29].

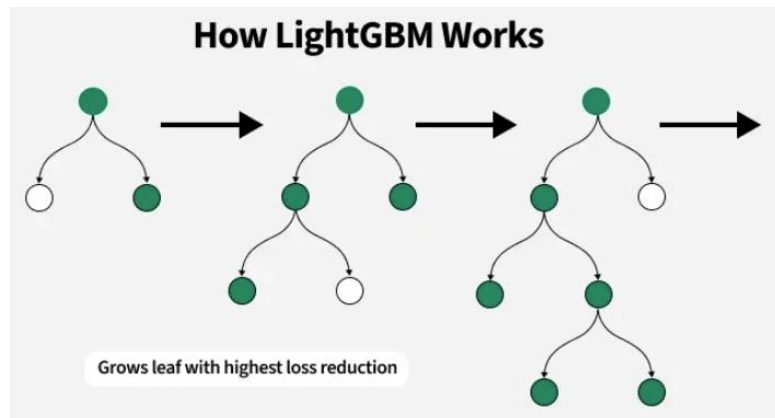


Рисунок 2.11 – Алгоритм роботи LightGBM [29]

Ще одна з реалізацій градієнтного бустингу, головною спеціалізацією якого є робота з категоріальними змінними є CatBoost (рис. 2.12). Основною перевагою вважається здійснення внутрішнього кодування класифікаційних ознак без виконання додаткових дій, тим самим процес препроцесингу даних значно спрощується, а також зменшується шанс виникнення інформаційного витоку. CatBoost оснащений інструментами регуляризації, та можливістю використання ресурсів графічного процесора, що позитивно відображується як на точності самих результатів так і на швидкості їх отримання [30].

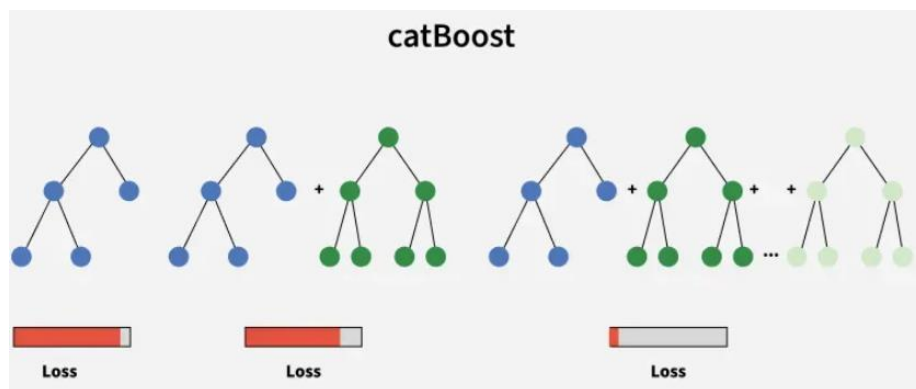


Рисунок 2.12 – Алгоритм роботи CatBoost [30]

Підсумовуючи, застосування принципу градієнтного бустингу, та трьох його описаних реалізацій, може дати суттєвий поштовх при створенні моделі прогнозування рівня проблемного використання Інтернету дітьми, поєднуючи високу точність результатів, можливість використання великих наборів різноманітних даних, зменшений ризик перенавчання та швидкість отримання фінальних передбачень [28; 29; 30].

## **2.4. Нейромережеві моделі**

Нейронні мережі – потужний інструмент машинного навчання, побудований за принципом роботи мозку людини. Їх можливості включають в себе проведення операцій з обробки даних та виявлення взаємозв'язків між ознаками [30].

Такі мережі зазвичай складаються з наступних частин:

1. Нейрони – ключові елементи нейромереж, що підпорядковуються завчасно визначеним порогам та функціям активації, та до яких на вхід приходять сигнали;
2. Зв'язки – з'єднання нейронів, що переправляють інформацію між ними, керовані вагами та правилами;
3. Ваги та зміщення – характеристики, що встановлюють ступінь і значущість зв'язків;
4. Функції активації – методи обробки та передачі інформації серед нейронних шарів;
5. Правила оновлення параметрів – способи корекції вагів та зміщень, для покращення точності отримуваних результатів.

Шари – блоки нейронів, що формують нейромережу та відповідають за її роботу. Існує три типи шарів – вхідний (отримання даних), прихований (основні обчислення) та вихідний (формування результатів) (рис. 2.13).

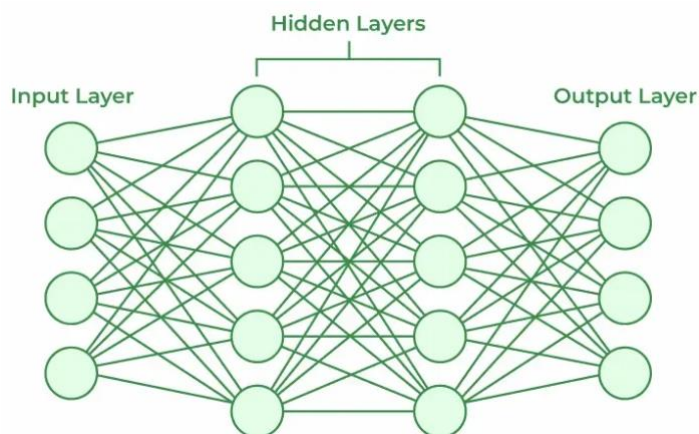


Рисунок 2.13 – Архтектура шарів нейронної мережі [31]

Процес навчання проходить у три головні етапи – пряме поширення сигналу, формування результатів на виході та ітеративна корекція параметрів для покращення якості прогнозів (рис. 2.14).

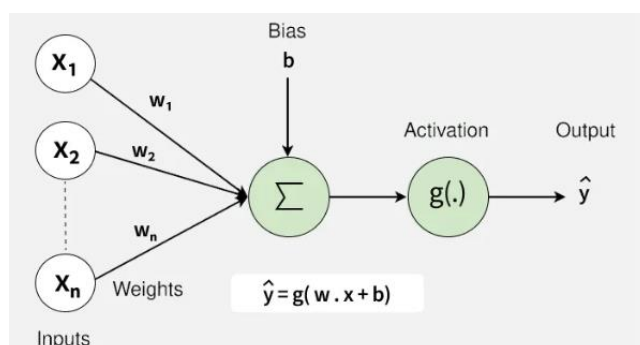


Рисунок 2.14 – Схема побудови нейронних мереж [31]

Загальна схема роботи алгоритму нейромереж [31]:

1. Пряме поширення – проходження даних безпосередньо зі вхідного шару, через приховані і до вихідного. В процесі, в кожному з нейронів відбувається лінійне перетворення вхідних даних (добуток вагів зі значеннями з додаванням зміщення) та проведення через функцію активації, додаючи нелінійність.

2. Зворотне поширення – оцінка продуктивності через визначення різниці між отриманими прогнозами та реальними даними. В процесі проводиться визначення похибки обчислень, далі градієнта за правилом

зворотнього поширення, та, на основі цього, зміна значень вагів (протилежний напрямок до напрямку градієнта) та зміщення.

3. Ітеративне повторення роботи нейромережі до моменту досягнення критерію зупинки її роботи.

Для прогнозування рівня проблемного використання Інтернету дітьми, нейромережеві моделі можуть стати в нагоді, з їх адаптивністю, можливістю виявлення глибоко прихованих нелінійних зв'язків у великих наборах даних та стратегіями паралельної обробки, що може значно зменшити час, необхідний для отримання результатів роботи.

## **2.5. Схема експериментів і валідації**

Важливою частиною дослідження є попереднє визначення схеми роботи з даними та експериментів, вибір метрик для оцінки якості та точності роботи алгоритмів.

### **2.5.1. Розбиття даних**

Розбиття даних – обов'язковий крок при роботі з алгоритмами машинного навчання, що передбачає розділення всього датасету на підвибірки для подальшої роботи [32].

Таким чином отримуються піднабір даних для навчання (навчальна вибірка) та піднабір для оцінки роботи фінальної моделі (тестувальна вибірка). Для валідації та налаштування конфігурації моделі може бути створена валідаційна підвибірка, або використовуватися крос-валідація з поділом навчального піднабору на рівні частини (при цьому розподіл класів в кожній з них є пропорційним). У такий спосіб гарантується незалежність та об'єктивність етапів роботи [32].

Основними кроками розбиття даних являються [32]:

1. Збір даних та їхній препроцесинг;

2. Розділення датасету на навчальну та тестувальну вибірки випадковим чином (зазвичай це співвідношення 70-80% до 20-30% залежно від кількості інформації);

3. Виділення частини даних для валідаційної вибірки (зазвичай 10-20%) або налаштування перехресної валідації;

4. Перевірки репрезентативності та узагальненості розподілення даних.

До розподілу даних слід підходити уважно, попередньо проаналізувавши наявні дані. Розмір вибірки, кількість і структура даних дослідження – все це може суттєво вплинути на результуючу схему роботи. Проте, незважаючи на складність вищеописаного процесу, він дозволяє уникати ризику перенавчання, сприяє полегшенню підбору гіперпараметрів алгоритмів та безпосередньо підвищує надійність передбачень моделі [32].

### 2.5.2. Метрика якості коефіцієнту згоди Каппа

Коефіцієнт Каппи Коена – статистична характеристика, що використовується для отримання картини якості роботи моделей машинного навчання за використання класифікаційних ознак. Цей показник показує рівень узгодженості між отримуваними результатами та реальними значеннями [33].

Коефіцієнт згоди Каппа знаходиться як відношення між різницею фактичної згоди та випадкової до максимальної можливої, з урахуванням випадковості експерименту (рис. 2.15).

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

Рисунок 2.15 – Формула обчислення Каппа коефіцієнту згоди [33]

Отримувані значення коефіцієнту згоди Каппа можна інтерпретувати за допомогою визначених меж [33]:

1. Менше 0 – відсутність узгодженості;
2. В межах  $[0; 0.2]$  – незначне узгодження;
3. В межах  $[0.21; 0.4]$  – довірче узгодження;
4. В межах  $[0.41; 0.6]$  – помірне узгодження;
5. В межах  $[0.61; 0.8]$  – значне узгодження;
6. В межах  $[0.81; 1]$  – (майже) ідеальне узгодження.

## **Висновки до розділу 2**

Цей розділ відобразив у собі всі засоби реалізації, що будуть використовуватися в практичній частині роботи, зокрема – різноманітні підходи та алгоритми машинного навчання, що стануть основними інструментами прогнозування рівня проблемного використання мережі Інтернет дітьми, використовуючи фізіологічні, статистичні та інші показники.

Такими стали базові методи (лінійна та логістична регресії, дерева рішень), поодинокі просунуті алгоритми (метод найближчих сусідів, машина опорних векторів), ансамблеві підходи (метод випадкового лісу, градієнтного бустингу та його реалізацій XGBoost, LightGBM, CatBoost), а також нейронні мережі.

Розглянуто схему експериментів та валідації результатів, що поєднує в собі розбиття даних, а також оцінку точності та узгодженості передбачень моделей на основі метрики якості коефіцієнту згоди Каппа.

Підсумовуючи, можна сказати, що здійснений аналіз надалі стане фундаментом створення повноцінного та комплексного підходу для досягнення поставленої мети дослідження.

### **РОЗДІЛ 3**

## **ОБҐРУНТУВАННЯ ВИБОРУ ТЕХНОЛОГІЧНОГО ТА МОДЕЛЬНОГО СТЕКА ДЛЯ СТВОРЕННЯ СИСТЕМИ ПРОГНОЗУВАННЯ РІВНЯ ПРОБЛЕМНОГО ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ**

Цей розділ містить опис початкової стадії практичної реалізації поставленої задачі прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників з практичної точки зору. Відбувається початковий огляд та вибір технологічного стека, за допомогою якого буде проводитися подальша робота. Також наведено обґрунтування та опис набору даних для наступних кроків, а також роботу над ним. Останнім пунктом є детальний опис стадії вибору фінальної моделі, яка застосовуватиметься в програмному застосунку, порівняння результатів роботи різноманітних моделей машинного навчання, нейронних мереж, їх оптимізованих версій, ансамблів, тим самим доводячи доцільність використання одного конкретного варіанту.

### **3.1. Обґрунтування вибору технологічного стека**

Ледь не жодне дослідження практичного характеру, в сфері машинного навчання, не обходиться без використання певного технологічного стека – зазвичай це певна мова програмування, її бібліотеки та фреймворки, середовище розробки. Саме тому, є логічним та важливим кроком, перед початком виконання практичних дій – обрати інструменти їх виконання. Правильний вибір може значно полегшити подальшу роботу, скоротити витрати часових та обчислювальних ресурсів.

### 3.2.1. Вибір мови програмування

Усі подальші дії не мали б сенсу без розуміння, яка мова програмування буде застосовуватися під час проведення дослідження. Тому першопочатково, варто зосередитись на цьому.

З кожним роком кількість популярних та різноманітних мов програмування на будь-який смак зростає. Всі вони постійно отримують оновлення та покращення, збагачують свій функціонал та набір можливостей. Незважаючи на це, кожна з мов програмування, хоча і може застосовуватися в різних сферах, негласно належить до вузького напрямку, проявляючи себе в ньому найкраще. Пристосованість під конкретні задачі – важливий фактор при виборі мови програмування.

Саме тому, для подальшої роботи над практичною реалізацією завдання з прогнозування, було розглянуто набір з чотирьох мов програмування, які можуть бути використані в задачах машинного навчання, проведено їх порівняння та обрано найбільш підходящу.

Таким чином, порівняльний аналіз різних мов програмування представлено нижче (рис. 3.1):

| Ознака              | Python                                                                                                                              | R                                                                                                     | Java                                                                                           | C++                                                                                                                          |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| Домінуючий напрям   | Методи машинного навчання, аналіз та робота з даними                                                                                | Методи статистичного аналізу, візуалізація результатів                                                | BigTech проекти                                                                                | Розробка систем з високою продуктивністю                                                                                     |
| Операційні системи  | Крос-платформений                                                                                                                   | Крос-платформений                                                                                     | Крос-платформений                                                                              | Крос-платформений                                                                                                            |
| Ефективність        | Середня ефективність, автоматизованість роботи з системними ресурсами                                                               | Нижче середньої ефективності, автоматизованість роботи з системними ресурсами                         | Висока ефективність, оптимізованість роботи з системними ресурсами за використанням JVM        | Висока ефективність, ручне керування системними ресурсами                                                                    |
| Зручність           | Простий для розуміння та інтерпретації, легкий в вивченні, зручний при використанні                                                 | Простий для розуміння та інтерпретації в сфері статистичного аналізу, складніший для інших задач      | Середня складність для розуміння та інтерпретації, складність через строгу типізацію даних     | Висока складність для розуміння та інтерпретації, запутаніший синтаксис, необхідність ручного керування ресурсами            |
| Універсальність     | Висока універсальність, найкращий вибір в сфері домінуючого напрямку                                                                | Висока універсальність, найкращий вибір в сфері домінуючого напрямку                                  | Висока універсальність, найкращий вибір в сфері домінуючого напрямку                           | Висока універсальність, найкращий вибір в сфері домінуючого напрямку                                                         |
| Бібліотеки          | Основний акцент на бібліотеках машинного навчання (TensorFlow, scikit-learn) бібліотеках аналізу та роботи з даними (NumPy, Pandas) | Основний акцент на бібліотеках для статистичного аналізу (lme4) та візуалізації результатів (ggplot2) | Основний акцент на бібліотеках розробки великих, високонавантажених систем (Hibernate, Spring) | Основний акцент на бібліотеках розробки складних математичних алгоритмів для роботи з великою кількістю даних (Eigen, Boost) |
| Середовища розробки | Google Colab, PyCharm, VS Code                                                                                                      | RStudio by Posit, PyCharm, JupyterLab                                                                 | Eclipse, IntelliJ IDEA, Apache NetBeans                                                        | Clion, Apache NetBeans, Visual Studio                                                                                        |

Рисунок 3.1 – Порівняльний аналіз мов програмування [34]

Для розробки прогнозувальної моделі машинного навчання в рамках поставленої задачі, можна зробити висновок, що найбільш оптимальним рішенням буде використання мови програмування Python.

### **3.2.2. Вибір бібліотек**

Після того, як було визначено мову програмування Python для побудови системи прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників, обов'язковим кроком є відбір бібліотек і фреймворків даної мови програмування, які використовуватимуться для створення, навчання та подальшого використання моделей машинного навчання та нейронних мереж, а також реалізації на основі фінальної моделі програмного застосунку з інтерфейсом користувача.

Основними факторами, за якими обирались відповідні фреймворки та бібліотеки стали швидкість та якість роботи, інтегрованість із засобами для роботи з даними та їх сумісністю між собою, гнучкість та придатність в контексті поставленої задачі.

Таким чином, було сформовано кілька категорій для різних етапів роботи, та визначено найкращі бібліотеки та фреймворки для кожної з них.

Бібліотеки для роботи з даними:

1. NumPy – бібліотека, що застосовується в роботі з багатовимірними масивами, векторизованими розрахунками [35].

2. Pandas – бібліотека, що застосовується для перенесення даних в оптимізовану, зручну для використання структуру даних, яка в подальшому використовується в більшості операцій при роботі з даними (імпорт та експорт, очищення, заповнення та відновлення, поєднання, трансформації) [35].

3. Tabulate – допоміжна бібліотека, що застосовується для гарного подання в текстовому вигляді результатів у форматі таблиці [35].

Бібліотеки машинного навчання:

1. Scikit-learn – головна бібліотека реалізацій методів та підходів до машинного навчання (алгоритмів, метрик оцінки, роботи з даними тощо) [35].

2. PyTorch – бібліотека глибокого навчання, що застосовується для створення та навчання нейронних мереж. Virізняється своєю можливістю використання динамічної обчислювальної графіки, прискоренням графічного процесора, широким спектром готових до використання компонентів [35].

3. Skorch – бібліотека, що містить в собі комбінацію переваг Scikit-learn та PyTorch, надаючи унікальний формат роботи методів навчання та передбачення, тим самим полегшуючи додавання нейромережових моделей у класичні процеси створення моделей машинного навчання, з можливістю використання кросвалідації [35].

4. XGBoost, LightGBM, CatBoost – реалізації конкретних алгоритмів градієнтного бустингу, які не ввійшли в бібліотеку Scikit-learn [35].

Інструменти для візуалізації результатів роботи з даними:

1. Matplotlib – основна бібліотека, що застосовується для візуального зображення графіків, що можуть вказувати на результати проміжних кроків роботи в задачі машинного навчання для їх кращої інтерпретації [35].

2. Seaborn – бібліотека, побудована на основі Matplotlib, що застосовується для візуального відображення статистичних даних, під час аналізу та роботи з датасетами та їх ознаками [35].

Фреймворки побудови інтерфейсу користувача та збереження моделей:

1. Tkinter – одна з найбільш використовуваних, зручних для розуміння споживачем бібліотек, які застосовується для створення графічного інтерфейсу користувача [35].

2. Joblib – бібліотека для збереження та подальшого завантаження моделей, препроцесорів, різноманітних гіперпараметрів у єдиному форматі для повторного застосування – зникає необхідність постійного навчання моделей тощо, що може займати великі проміжки часу [35].

В результаті, було сформовано обширний набір бібліотек, інструментів та фреймворків – все необхідне для практичної реалізації завдання мовою програмування Python якнайшвидше та найбільш зручно.

### 3.2.3. Вибір середовища розробки

Останнім, але не менш важливим кроком у виборі технологічного стека є вибір середовища розробки, яке буде зручним у використанні, а також міститиме багатий функціонал при роботі з мовою програмування Python, та всіма необхідними бібліотеками.

Для цього було розглянуто три основні середовища розробки, що застосовуються разом із вибраною мовою програмування – Google Colaboratory, PyCharm та Visual Studio Code (рис. 3.2).

| Критерій                            | Google Colab                                                                                                       | PyCharm                                                                                                  | VS Code                                                                           |
|-------------------------------------|--------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Вид середовища                      | Браузерне середовище розробки                                                                                      | Інтегроване середовище розробки                                                                          | Code Editor з додатковими можливостями (модулями)                                 |
| Необхідність покупки платної версії | Безкоштовна версія (можливості збереження даних та виконання блоків коду обмежені)                                 | Безкоштовна версія (Community)                                                                           | Безкоштовна версія                                                                |
| Можливість локальної роботи         | Локальне тестування обмежене. Виконання блоків коду в хмарному середовищі                                          | Локальна розробка та тестування з розширеними можливостями                                               | Локальна розробка та тестування                                                   |
| Функціонал                          | Обмежений інструментарій, акцент на виконання коду                                                                 | Широкий інструментарій для розробки та тестування (можливість покрокового проходження, код рефакторингу) | Вузкий інструментарій, розширюється за допомогою встановлення додаткових модулів  |
| Використання ресурсів               | Використовуються зовнішні ресурси, неможливість виконання за відсутності з'єднання з мережею та зовнішніх ресурсів | Використовуються ресурси системи, висока продуктивність, оптимізованість роботи                          | Використовуються ресурси системи, встановлені модулі потребують додаткових витрат |

Рисунок 3.2 – Порівняльний аналіз середовищ розробки [36]

Для розробки прогнозувальної моделі машинного навчання та програмного застосунку в рамках поставленої задачі, було обрано середовище PyCharm, зrt надає всі необхідні інструменти для досягнення мети дослідження.

## 3.2. Обґрунтування вибору та робота з набором даних

Вибір датасету, аналіз, інтерпретація та обробка даних – це все невід'ємні частини роботи над проєктом у сфері машинного навчання. Саме

тому, значну увагу перед початком практичної реалізації конкретної моделі варто приділити роботі з набором даних.

### **3.2.1. Загальний огляд датасету**

Першопочатково, після вибору тематики дослідження, було звернено увагу на набір даних із відкритого змагання під назвою «Problematic Internet Use» від Child Mind Institute, яке полягало в знаходженні зв'язку між фізичними показниками й активністю дитини та проблемним використанням Інтернету. Цей датасет являє собою незалежну частину проєкту Healthy Brain Network, та містить в собі до п'яти тисяч досліджуваних дітей та підлітків, віком від п'яти до двадцяти років. Кожен з них пройшов наукові та клінічні обстеження, що доповнюють один одне, роблячи його ідеальним для обраної теми дослідження. Основна мета даного проєкту – знайти певні біологічні маркери, що зможуть зробити значний поштовх у сфері діагностики та лікування як психічних, так і когнітивних розладів, спираючись на сучасні технології, та маючи наукове обґрунтування [19; 37].

Повертаючись до змагання, провівши первинний огляд набору, можна побачити, що він містить два основних розділи даних – дані фізичної активності, зокрема акселерометричні записи та результати різних фіз-тестів, а також поведінкові дані – опитувальники, пов'язані з використанням Інтернету ознаки тощо. Разом ми отримуємо потужний датасет, що ідеально підходить до поставленої задачі прогнозування [37].

Ще одним важливим моментом є спосіб зберігання даних, що в подальшому впливатиме на роботу з ними – акселерометричні дані представлені в форматі parquet-файлів, і є унікальними для кожного з учасників. Табличні дані – одна спільна таблиця в CSV-форматі [37].

Ключовими можна назвати дві характеристики – тест на Інтернет-залежність, заповнений батьками та їх дітьми – Parent-Child Internet Addiction Test (PCIAT), а також побудований на його основі (використовуючи завчасно визначені межі) стандартизований показник, що напряду вказує на проблемне використання Інтернету – Severity Impairment Index (SII) [19; 37; 38].

Негативною особливістю даного набору даних є наявність значної кількості пропущених значень (навіть серед цільових змінних), що створює необхідність в попередній обробці та заповненні даних. Водночас, це гарна можливість протестувати широкий спектр алгоритмів, методик та підходів до роботи з різноманітними даними [37].

### 3.2.2. Робота з даними актиграфії

Першим етапом роботи з даними було обрано обробку актиграфічних даних, зібраних за допомогою носимих наручних сенсорів, в основі яких встановлено акселерометр. Саме вони слугують основним джерелом даних щодо фізичної активності, а також біоповедінкових особливостей дітей, що беруть участь у дослідженні. Всі записи – це часові ряди з інтервалів по п'ять секунд [37].

Після зчитування всіх даних актиграфії та їх огляду, можна створити описову таблицю, зображену на рисунку нижче (рис. 3.3).

| Змінна              | Опис                                                                                                         |
|---------------------|--------------------------------------------------------------------------------------------------------------|
| id                  | Ідентифікатор пацієнта, який відповідає полю id у train/test.csv.                                            |
| step                | Цілий часовий крок для кожного спостереження в серії.                                                        |
| X, Y, Z             | Прискорення (в g), яке фіксує наручний акселерометр уздовж кожної з осей.                                    |
| enmo                | Євклідова норма мінус один (ENMO) для сигналів акселерометра (X, Y, Z), від'ємні значення округлені до нуля. |
| anglez              | Кут руки відносно горизонтальної площини, обчислений з компонентів акселерометра.                            |
| non-wear_flag       | 0 = годинник одягнений, 1 = годинник знятий (за визначенням GGIR).                                           |
| light               | Вимірювання рівня освітлення (люкси).                                                                        |
| battery_voltage     | Напруга батареї у мілівольтах (mV).                                                                          |
| time_of_day         | Час початку 5-секундного інтервалу вимірювання (%H:%M:%S.%9f).                                               |
| weekday             | День тижня (1 = понеділок, ..., 7 = неділя).                                                                 |
| quarter             | Квартал року (від 1 до 4).                                                                                   |
| relative_date_PCiAT | Кількість днів відносно тесту PCiAT (від'ємні значення = до тесту).                                          |

Рисунок 3.3 – Описова таблиця актиграфічних даних

Розглянемо наявні дані більш детально. Перша ознака, на яку варто звернути увагу – величина, якою характеризується інтенсивність рухів

піддослідного – Euclidean Norm Minus One (ENMO). Саме вона визначає поточну категорію активності людини (від сидячого стану до високої фізичної активності). В парі з ним не варто упускати з уваги показник кута нахилу відносно горизонтальної поверхні в просторі – Angle-Z. Всі інші характеристики на перший погляд не є ключовими, але можуть містити приховані зв'язки, тому будуть використовуватися для подальшої роботи [37].

Обробку даних варто розпочати зі зміни частоти часових рядів, нормалізувавши її до однієї хвилини. Таким чином, зменшується загальний обсяг інформації, вирівнюються дані за часом та частково позбуваються пропущені значення. Наступним кроком є заповнення короткочасних пропущених вимірювань (яскравим прикладом яких є відсутність сигналу за один, або декілька інтервалів) та усунення зашумленості.

Важливим є визначення порогів фізичної активності, для їх класифікації за рівнем інтенсивності рухової активності. В якості основного методу було обрано модель суміші нормальних (гаусових) розподілів, а альтернативним – метод, в основі якого лежить використання квантилів. Таким чином, всі часові інтервали були поділені на три основних рівні активності – сидячий (або ж дуже низький), легкий та помірний (або ж інтенсивний) (рис. 3.4).

```
Середні GMM: [0.008700000122189522, 0.046900000042319298, 0.12229999899864197]
Пороги підібрані: sedentary_upper=0.0191, mvpa_lower=0.1058
```

Рисунок 3.4 – Визначення порогів активності

Подальша обробка полягала у визначенні:

1. Денних метрик активності, а саме середніх та перцентильних значень ENMO, кількісного значення хвилин залежно від рівня активності, години, впродовж якої відбувається найбільша активність, варіативності певних показників.

2. Ознак, які пов'язані з циркадними ритмами та сном, а саме значень кількості та якості сну (зокрема інформації щодо пробуджень), характеристиками циркадного ритму.

Внаслідок таких дій, для кожного з учасників (дані щодо яких наявні) вибудовується спільний набір характеристик та ознак, що містить в собі повний спектр інформації по даним актиграфії, без включення неякісних та непотрібних даних. Такий набір зберігається в CSV-формат, для подальшої інтеграції та роботи з іншими наявними табличними даними.

Для узагальнення проведеної роботи, та отримання цілісної картини щодо поточного розділу даних, корисним буде побудова статистик ознак. Сильне занепокоєння викликає кількість годин, проведених уві сні – вона являється аномально низькою (рис. 3.5).

| variable               | mean    | std     | min    | 5%     | 25%    | 50%    | 75%    | 95%    | max    | missing_frac |
|------------------------|---------|---------|--------|--------|--------|--------|--------|--------|--------|--------------|
| valid_intervals_mean   | 888.313 | 235.886 | 389    | 719    | 723    | 780    | 975    | 1440   | 1440   | 0            |
| enmo_mean_mean         | 0.0479  | 0.0259  | 0.0003 | 0.0118 | 0.0282 | 0.044  | 0.0636 | 0.0959 | 0.1785 | 0            |
| enmo_p95_mean          | 0.1062  | 0.0575  | 0.0011 | 0.0318 | 0.0659 | 0.0991 | 0.1335 | 0.2087 | 0.4774 | 0            |
| minutes_sedentary_mean | 262.676 | 263.652 | 0      | 0      | 12.75  | 175    | 486.25 | 705.75 | 1343   | 0            |
| minutes_light_mean     | 541.108 | 213.457 | 0      | 106.75 | 420.5  | 595    | 690    | 827.25 | 1106   | 0            |
| minutes_mvpa_mean      | 68.51   | 90.9586 | 0      | 0      | 4      | 32     | 97     | 264.25 | 633    | 0            |
| mvpa_share_mean        | 0.0826  | 0.1123  | 0      | 0      | 0.0043 | 0.0387 | 0.117  | 0.3145 | 0.7819 | 0            |
| sleep_minutes_mean     | 59.9849 | 99.5744 | 0      | 0      | 0      | 1      | 74.25  | 287.5  | 360    | 0            |
| awakenings_mean        | 5.2149  | 9.4394  | 0      | 0      | 0      | 0      | 6      | 28     | 52     | 0            |
| rhythm_amplitude_mean  | 0.0874  | 0.0512  | 0.0021 | 0.0225 | 0.0517 | 0.0797 | 0.1122 | 0.1801 | 0.5996 | 0            |
| rhythm_stability_mean  | 0.0256  | 0.0142  | 0.0007 | 0.0061 | 0.0148 | 0.024  | 0.0331 | 0.0509 | 0.1361 | 0            |
| acrophase_mode         | 14.002  | 3.7508  | 0      | 8      | 12     | 14     | 16     | 20     | 23     | 0            |

Рисунок 3.5 – Статистики за актиграфічними даними

Окрім розрахунку статистичних показників, необхідним є проведення візуалізації найбільш показових і важливих даних на графіках для кращого розуміння.

У сидячій поведінці дітей спостерігається досить висока варіативність (починаючи від нульових результатів, і доходючи до двадцятигодинних періодів) (рис. 3.6).

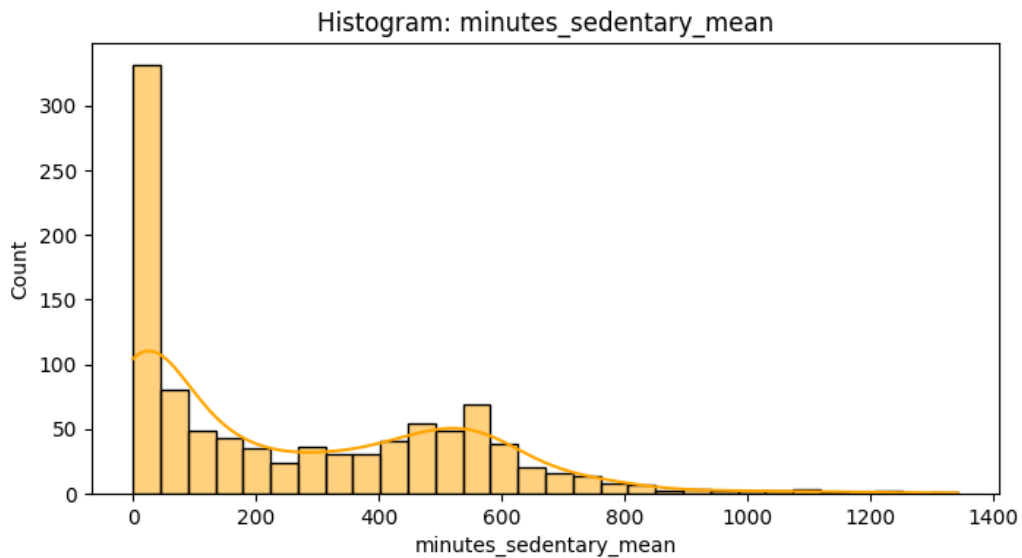


Рисунок 3.6 – Гістограма «Розподіл часу сидячої поведінки»

Як можна побачити, у більшості досліджуваних дітей наявна легка активність впродовж значного періоду часу (десятигодинна медіана) (рис. 3.7).

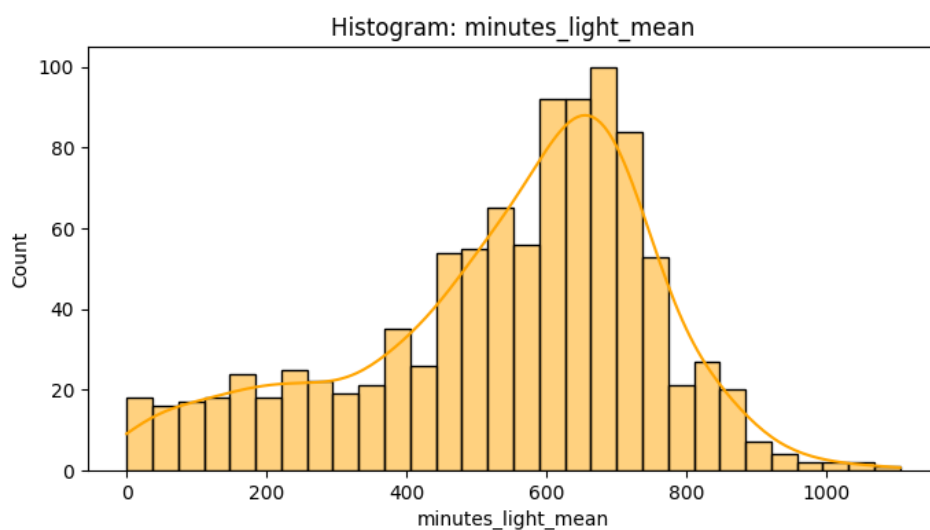


Рисунок 3.7 – Гістограма «Розподіл часу легкої активності»

Натомість помірна та висока активність, майже відсутні (тридцятихвилинна медіана) (рис. 3.8).

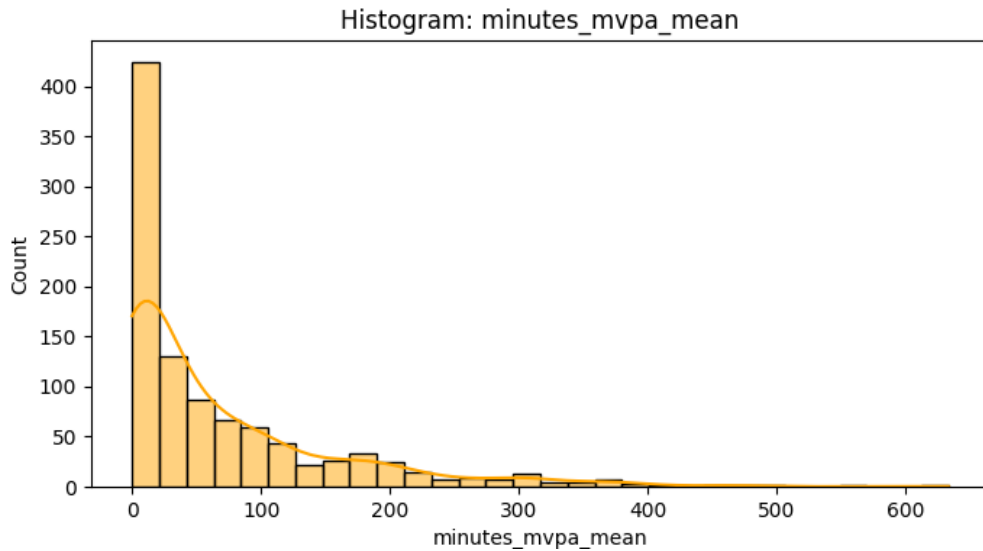


Рисунок 3.8 – Гістограма «Розподіл часу помірної/високої активності»

Сон проходить у більшості випадків спокійно, без пробуджень (за винятком невеликої кількості дітей) (рис. 3.9).

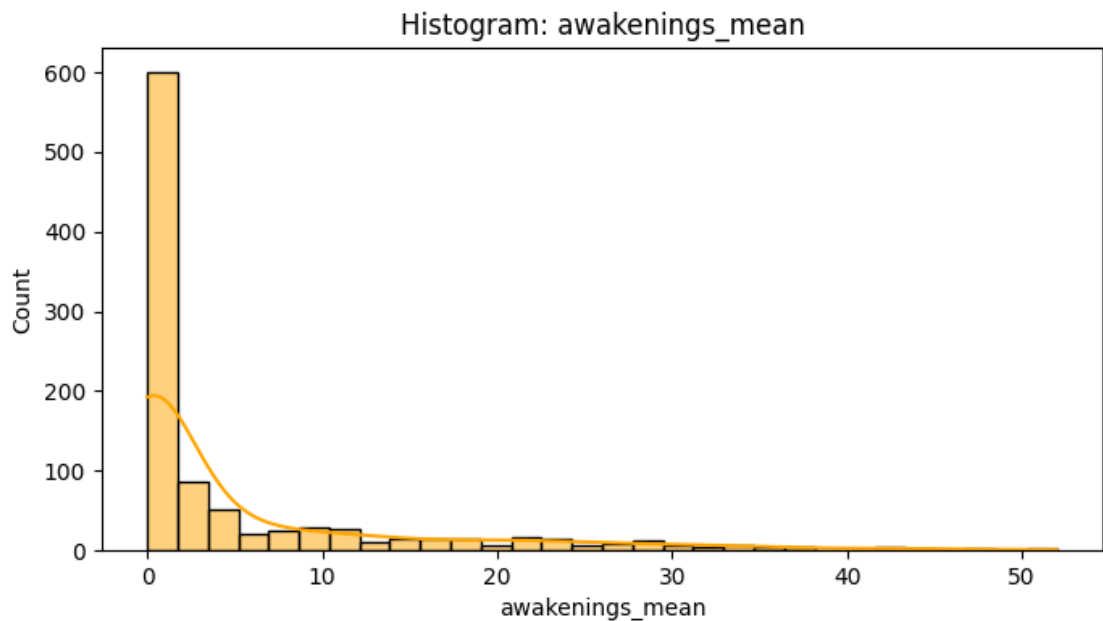


Рисунок 3.9 – Гістограма «Кількість нічних пробуджень»

Циркадні характеристики вказують на виражений денний ритм у дітей – висока амплітуда (рис. 3.10) та низька стабільність (рис. 3.11).

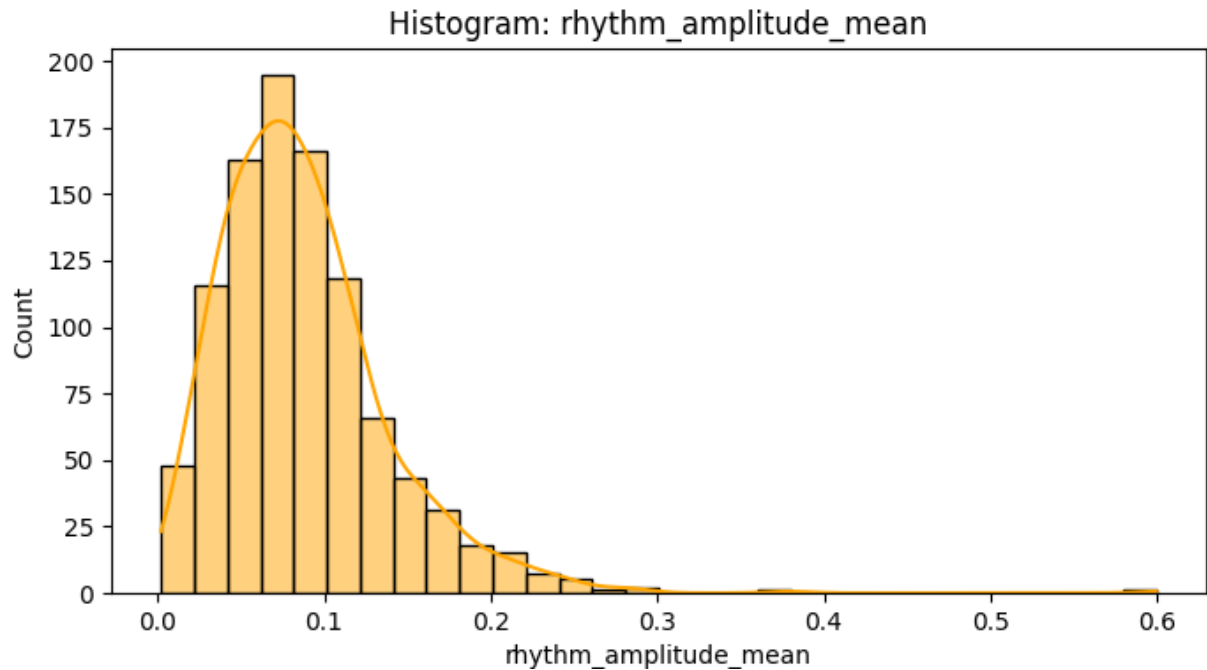


Рисунок 3.10 – Гістограма «Амплітуда циркадних ритмів»

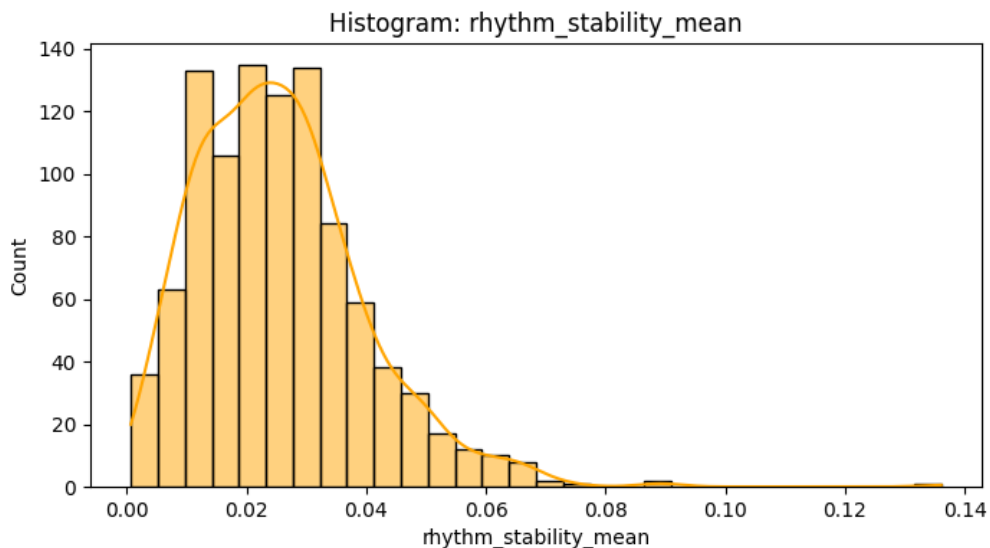


Рисунок 3.11 – Гістограма «Стабільність циркадних ритмів»

Також, гарно видно, що годинами пікової активності дітей є години від 12 до 16 (рис. 3.12).

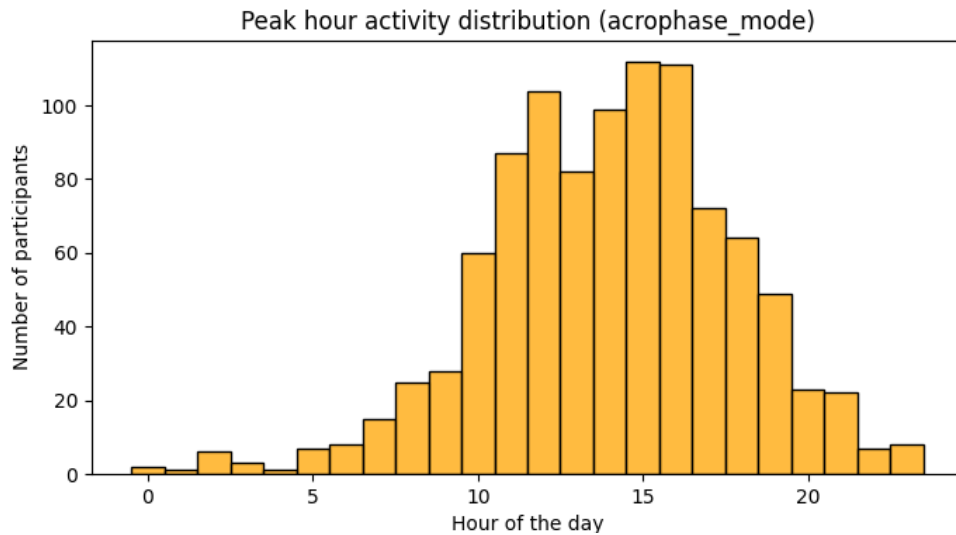


Рисунок 3.12 – Гістограма «Години пікової активності протягом доби»

Scatter-графіки не дали чіткого розуміння залежності між кількістю годин сну та середньодобовою активністю, оскільки показники тривалості сну являються стабільно низькими для всіх (рис. 3.13).

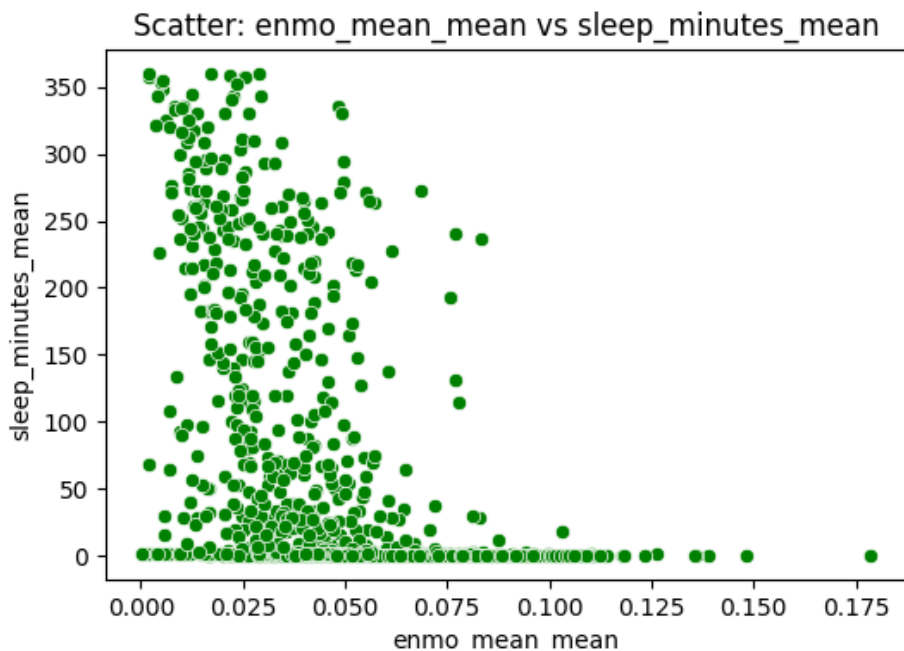


Рисунок 3.13 – Scatter-графік «Залежність середньої активності від сну»

У той самий час простежується досить сильна лінійна залежність абсолютного й відносного часу інтенсивної активності (рис. 3.14).

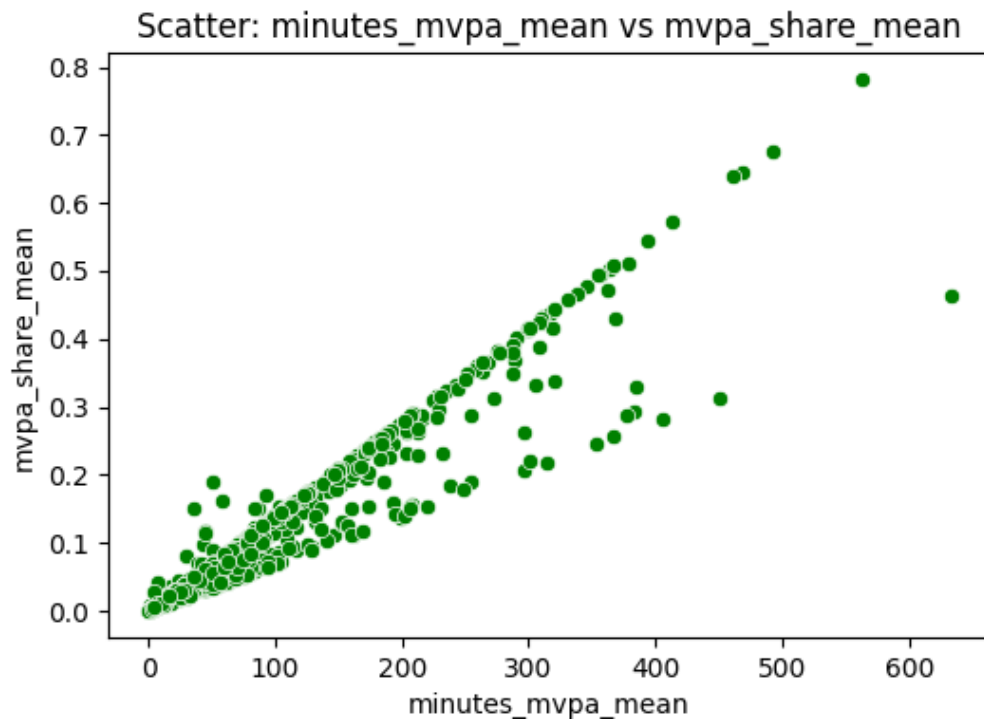


Рисунок 3.14 – Scatter-графік «Співвідношення абсолютного й відносного часу інтенсивної активності»

Підсумовуючи, маємо такі результати: більшість дітей мають низький рівень активності та тривалості сну (хоча сон і є досить спокійним), варіативну сидячу поведінку та лінійну залежність між абсолютним і відносним часом інтенсивної активності.

### 3.2.3. Робота з табличними даними

Другою половиною датасету є таблична частина даних з результатами різноманітних вимірювань та анкетних опитувальників. Вони були сформовані різноманітними інструментами від Healthy Brain Network, та включають в себе фізичні параметри, статистичні ознаки, результати психологічних анкет та опитувальників, інші характеристики. Поєднавши всі, можна виділити наступні основні категорії [37]:

1. Демографічні ознаки;
2. Дані про тривалість використання мережі Інтернет, пристроїв загалом;
3. Дані по загальному функціонуванню дітей за шкалою CGAS;
4. Фізіологічні ознаки;

5. Кардіореспіраторна витривалість згідно з протоколом NHANES;
6. Загальна фізична підготовка;
7. Склад тіла та похідні ознаки;
8. Дані про фізичну активності за останній тиждень;
9. Оцінки порушень пов'язаних зі сном;
10. Опитувальник PCİAT, що складається з двадцяти пунктів та сформований на його основі індекс Severity Impairment Index (значення якого варіюються від 0 (None) до 3 (Severe)) [38].

Зчитавши табличну частину датасету з Excel-файлу, а також її огляд, було створено описову таблицю, за допомогою якої можна ознайомитися зі всіма наявними табличними даними.

Таким чином, перша частина табличних даних наведено на рисунку нижче (рис. 3.15).

| Variable                     | Description                           |
|------------------------------|---------------------------------------|
| id                           | Ідентифікатор пацієнта                |
| Basic_Demos-Enroll_Season    | Сезон зарахування                     |
| Basic_Demos-Age              | Вік учасника                          |
| Basic_Demos-Sex              | Стать учасника                        |
| C6AS-Season                  | Сезон участі                          |
| C6AS-C6AS_Score              | Бал за шкалою загальної оцінки дитини |
| Physical-Season              | Сезон участі                          |
| Physical-BMI                 | Індекс маси тіла (кг/м <sup>2</sup> ) |
| Physical-Height              | Зріст (дюйми)                         |
| Physical-Weight              | Вага (фунти)                          |
| Physical-Waist_Circumference | Окружність талії (дюйми)              |
| Physical-Diastolic_BP        | Діастолічний тиск (мм рт. ст.)        |
| Physical-HeartRate           | Частота серцевих скорочень (уд/хв)    |

Рисунок 3.15 – Описова таблиця табличних даних (1 частина)

Наступна, друга частина табличних даних наведено на рисунку нижче (рис. 3.16).

|                             |                                            |
|-----------------------------|--------------------------------------------|
| Physical-Systolic_BP        | Систолічний тиск (мм рт. ст.)              |
| Fitness_Endurance-Season    | Сезон участі                               |
| Fitness_Endurance-Max_Stage | Максимальний досягнутий етап               |
| Fitness_Endurance-Time_Mins | Точний час виконання: хвилини              |
| Fitness_Endurance-Time_Sec  | Точний час виконання: секунди              |
| F6C-Season                  | Сезон участі                               |
| F6C-F6C_CU                  | Загальна кількість підйомів тулуба         |
| F6C-F6C_CU_Zone             | Фітнес-зона підйомів тулуба                |
| F6C-F6C_GSND                | Загальна сила хвату (недомінантна рука)    |
| F6C-F6C_GSND_Zone           | Фітнес-зона сили хвату (недомінантна рука) |
| F6C-F6C_GSD                 | Загальна сила хвату (домінантна рука)      |
| F6C-F6C_GSD_Zone            | Фітнес-зона сили хвату (домінантна рука)   |
| F6C-F6C_PU                  | Загальна кількість віджимань               |
| F6C-F6C_PU_Zone             | Фітнес-зона віджимань                      |

Рисунок 3.16 – Описова таблиця табличних даних (2 частина)

Наступна, третя частина табличних даних наведено на рисунку нижче (рис. 3.17).

|                            |                                                  |
|----------------------------|--------------------------------------------------|
| F6C-F6C_SRL                | Загальний показник нахилу вперед (ліва сторона)  |
| F6C-F6C_SRL_Zone           | Фітнес-зона нахилу вперед (ліва сторона)         |
| F6C-F6C_SRR                | Загальний показник нахилу вперед (права сторона) |
| F6C-F6C_SRR_Zone           | Фітнес-зона нахилу вперед (права сторона)        |
| F6C-F6C_TL                 | Загальний показник підйому тулуба                |
| F6C-F6C_TL_Zone            | Фітнес-зона підйому тулуба                       |
| BIA-Season                 | Сезон участі                                     |
| BIA-BIA_Activity_Level_num | Рівень активності                                |
| BIA-BIA_BMC                | Мінеральний вміст кісткової тканини              |
| BIA-BIA_BMI                | Індекс маси тіла                                 |
| BIA-BIA_BMR                | Базальний метаболізм                             |
| BIA-BIA_DEE                | Добові енергетичні витрати                       |
| BIA-BIA_ECW                | Позаклітинна вода                                |
| BIA-BIA_FFM                | Безжирова маса тіла                              |
| BIA-BIA_FFMI               | Індекс безжирової маси                           |

Рисунок 3.17 – Описова таблиця табличних даних (3 частина)

Наступна, четверта частина табличних даних наведено на рисунку нижче (рис. 3.18).

|                   |                                                                                             |
|-------------------|---------------------------------------------------------------------------------------------|
| BIA-BIA_FMI       | Індекс жирової маси                                                                         |
| BIA-BIA_Fat       | Відсоток жиру в тілі                                                                        |
| BIA-BIA_Frame_num | Тип тілобудови                                                                              |
| BIA-BIA_ICW       | Внутрішньоклітинна вода                                                                     |
| BIA-BIA_LDM       | Суха безжирова маса                                                                         |
| BIA-BIA_LST       | М'язова тканина без жиру                                                                    |
| BIA-BIA_SMM       | Скелетна м'язова маса                                                                       |
| BIA-BIA_TBW       | Загальна вода в організмі                                                                   |
| PAQ_A-Season      | Сезон участі                                                                                |
| PAQ_A-PAQ_A_Total | Підсумковий показник активності (підлітки)                                                  |
| PAQ_C-Season      | Сезон участі                                                                                |
| PAQ_C-PAQ_C_Total | Підсумковий показник активності (діти)                                                      |
| PCIAT-Season      | Сезон участі                                                                                |
| PCIAT-PCIAT_01    | Як часто ваша дитина порушує встановлені вами часові обмеження для користування інтернетом? |
| PCIAT-PCIAT_02    | Як часто ваша дитина нехтує домашніми обов'язками, щоб більше часу проводити онлайн?        |

Рисунок 3.18 – Описова таблиця табличних даних (4 частина)

Наступна, п'ята частина табличних даних наведено на рисунку нижче (рис. 3.19).

|                |                                                                                                    |
|----------------|----------------------------------------------------------------------------------------------------|
| PCIAT-PCIAT_03 | Як часто ваша дитина віддає перевагу онлайн-часу замість спілкування з родиною?                    |
| PCIAT-PCIAT_04 | Як часто ваша дитина заводить нові знайомства серед користувачів інтернету?                        |
| PCIAT-PCIAT_05 | Як часто ви скаржитеся на кількість часу, яку ваша дитина проводить онлайн?                        |
| PCIAT-PCIAT_06 | Як часто оцінки вашої дитини погіршуються через час, проведений онлайн?                            |
| PCIAT-PCIAT_07 | Як часто ваша дитина перевіряє електронну пошту перед тим, як зробити щось інше?                   |
| PCIAT-PCIAT_08 | Як часто ваша дитина виглядає відстороненою від інших після знайомства з інтернетом?               |
| PCIAT-PCIAT_09 | Як часто ваша дитина стає оборонною чи потайною, коли її запитують, що вона робить онлайн?         |
| PCIAT-PCIAT_10 | Як часто ви ловите свою дитину на тому, що вона заходить в інтернет всупереч вашим заборонам?      |
| PCIAT-PCIAT_11 | Як часто ваша дитина проводить час наодинці у своїй кімнаті, граючи на комп'ютері?                 |
| PCIAT-PCIAT_12 | Як часто ваша дитина отримує дивні телефонні дзвінки від нових «онлайн-друзів»?                    |
| PCIAT-PCIAT_13 | Як часто ваша дитина кричить, зривається чи дратується, якщо її відволікають від інтернету?        |
| PCIAT-PCIAT_14 | Як часто ваша дитина здається більш втомленою і виснаженою, ніж до появи інтернету?                |
| PCIAT-PCIAT_15 | Як часто ваша дитина, перебуваючи офлайн, думає лише про повернення в інтернет?                    |
| PCIAT-PCIAT_16 | Як часто ваша дитина влаштовує істерики через ваші обмеження часу перебування онлайн?              |
| PCIAT-PCIAT_17 | Як часто ваша дитина обирає інтернет замість колишніх улюблених хобі чи занять на свіжому повітрі? |

Рисунок 3.19 – Описова таблиця табличних даних (5 частина)

Остання, шоста частина табличних даних наведено на рисунку нижче (рис. 3.20).

|                                        |                                                                                                                                |
|----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| PCIAI-PCIAI_18                         | Як часто ваша дитина злиться або стає агресивною, коли ви обмежуєте їй час перебування онлайн?                                 |
| PCIAI-PCIAI_19                         | Як часто ваша дитина обирає інтернет замість прогулянок із друзями?                                                            |
| PCIAI-PCIAI_20                         | Як часто ваша дитина почувається пригніченою, нервовою чи засмученою офлайн, але це зникає, коли вона повертається в інтернет? |
| PCIAI-PCIAI_Total                      | Загальний бал                                                                                                                  |
| SDS-Season                             | Сезон участі                                                                                                                   |
| SDS-SDS_Total_Raw                      | Сирий підсумковий бал                                                                                                          |
| SDS-SDS_Total_T                        | T-бал                                                                                                                          |
| PreInt_EduHx-Season                    | Сезон участі                                                                                                                   |
| PreInt_EduHx-computerinternet_hoursday | Кількість годин на день за комп'ютером/в інтернеті                                                                             |

Рисунок 3.20 – Описова таблиця табличних даних (6 частина)

Почнемо з числових змінних. Проаналізувавши всі дані та їх опис, можна дійти до висновку, що від'ємні значення не являються нормою, тому на поточному етапі вони замінюються пустими значеннями. Фізичні показники («Physical-»), з нульовими значеннями теж являються некоректними, і для чистоти експерименту також замінюються пропусками.

Також варто звернути увагу на викиди, так як вони теж можуть спотворити якість подальшої роботи. Тому будь-яке значення, яке перевищує п'ять стандартних відхилень від середнього значення так само змінюється пустою коміркою.

Далі поговоримо про конкретні категорії даних. Одним з таких є «Sit and Reach» тестування. Взявши середнє арифметичне між значеннями вимірів правої та лівої ноги, отримуємо нову ознаку, а також можливість видалити дві залежні попередні.

Для класифікації до певного відрізка часу денної активності, було створено нову ознаку, яка позначатиме, що дитина виконувала інтенсивну фізичну активність більш ніж 60 хвилин («MVPA\_60»).

Деякі ознаки майже з високою точністю повторюють інші. Таким чином, варто прибрати змінну «PAQ-A», лишивши в наборі змінних її більш точний аналог «PAQ-C».

За тим самим принципом прибираємо стандартизовані значення ознаки «T-Score», лишивши сирі показники шкали SDS «Raw». В такий спосіб

кількість вхідних ознак скорочується, що в подальшому може значно спростити практичне виконання поставленої задачі.

Показники сили хвату видаляємо через їхню валідність лише для певних вікових груп, які у датасеті не охоплені належним чином, відповідно ознака не є повноцінною.

Підсумовуючи, ми отримуємо фінальний, доповнений та «чистий» набір табличних ознак, який є готовим для наступних кроків практичної реалізації нашої роботи.

Для узагальнення проведеної роботи, та отримання цілісної картини щодо поточного розділу даних, корисним буде побудова статистик ознак, їх візуалізація на графіках.

Таким чином, перша частина статистичних показників табличних даних наведена на рисунку нижче (рис. 3.21).

| variable                     | count | mean    | std     | min     | 5%      | 25%     | 50%     | 75%     | 95%     | max     | missing_frac |
|------------------------------|-------|---------|---------|---------|---------|---------|---------|---------|---------|---------|--------------|
| Basic_Demos-Enroll_Season    | 3168  | 1.45739 | 1.15328 | 0       | 0       | 0       | 1       | 3       | 3       | 3       | 0            |
| Basic_Demos-Age              | 3168  | 10.4302 | 3.57148 | 5       | 6       | 8       | 10      | 13      | 17      | 22      | 0            |
| Basic_Demos-Sex              | 3168  | 0.37689 | 0.48468 | 0       | 0       | 0       | 0       | 1       | 1       | 1       | 0            |
| C6AS-Season                  | 2065  | 1.41937 | 1.10689 | 0       | 0       | 0       | 1       | 2       | 3       | 3       | 0.34817      |
| C6AS-C6AS_Score              | 1950  | 65.0456 | 11.7919 | 25      | 45      | 59      | 65      | 75      | 85      | 95      | 0.38447      |
| Physical-Season              | 1891  | 2.00635 | 0.81712 | 1       | 1       | 1       | 2       | 3       | 3       | 3       | 0.40309      |
| Physical-BMI                 | 2391  | 19.2605 | 4.75883 | 8.52244 | 14.2578 | 15.9139 | 17.8891 | 21.3812 | 28.8984 | 43.3377 | 0.24527      |
| Physical-Height              | 2404  | 55.9122 | 7.49675 | 33      | 45      | 50      | 55      | 62      | 69      | 78.5    | 0.24116      |
| Physical-Weight              | 2397  | 90.5072 | 42.8305 | 31.8    | 44.16   | 58.2    | 77.6    | 114.2   | 172.44  | 306.4   | 0.24337      |
| Physical-Waist_Circumference | 695   | 27.1806 | 5.50006 | 18      | 21      | 23      | 26      | 30      | 38      | 50      | 0.78062      |
| Physical-Diastolic_BP        | 2343  | 69.5911 | 13.2562 | 14      | 52      | 61      | 68      | 76      | 94      | 136     | 0.26042      |
| Physical-HeartRate           | 2357  | 81.6037 | 13.6269 | 27      | 61      | 72      | 81      | 90      | 104     | 138     | 0.256        |
| Physical-Systolic_BP         | 2345  | 117.011 | 16.7924 | 49      | 95      | 107     | 115     | 125     | 149     | 198     | 0.25979      |
| Fitness_Endurance-Season     | 1060  | 1.45472 | 1.16075 | 0       | 0       | 0       | 2       | 2       | 3       | 3       | 0.6654       |
| Fitness_Endurance-Max_Stage  | 607   | 4.88303 | 1.66734 | 0       | 2       | 4       | 5       | 6       | 7       | 12      | 0.8084       |
| F6C-Season                   | 2674  | 1.38257 | 1.1347  | 0       | 0       | 0       | 1       | 2       | 3       | 3       | 0.15593      |

Рисунок 3.21 – Статистики за табличними даними (1 частина)

Наступна, друга частина статистичних показників табличних даних наведена на рисунку нижче (рис. 3.22).

|                            |      |         |         |         |         |         |         |         |         |         |         |
|----------------------------|------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| F6C-F6C_CU                 | 1846 | 10.9149 | 10.7428 | 0       | 0       | 3       | 9       | 15      | 31.75   | 70      | 0.4173  |
| F6C-F6C_CU_Zone            | 1822 | 0.47585 | 0.49955 | 0       | 0       | 0       | 0       | 1       | 1       | 1       | 0.42487 |
| F6C-F6C_PU                 | 1842 | 5.4506  | 6.9452  | 0       | 0       | 0       | 3       | 9       | 20      | 40      | 0.41856 |
| F6C-F6C_PU_Zone            | 1814 | 0.33241 | 0.47121 | 0       | 0       | 0       | 0       | 1       | 1       | 1       | 0.4274  |
| F6C-F6C_TL                 | 1855 | 9.23906 | 3.00746 | 0       | 4       | 7       | 10      | 12      | 13      | 22      | 0.41446 |
| F6C-F6C_TL_Zone            | 1823 | 0.78826 | 0.40865 | 0       | 0       | 1       | 1       | 1       | 1       | 1       | 0.42456 |
| BIA-Season                 | 1712 | 1.39836 | 1.05307 | 0       | 0       | 1       | 1       | 2       | 3       | 3       | 0.4596  |
| BIA-BIA_Activity_Level_num | 1593 | 2.67797 | 1.03573 | 1       | 1       | 2       | 3       | 3       | 4       | 5       | 0.49716 |
| BIA-BIA_BMC                | 1567 | 4.50848 | 2.30185 | 0.00989 | 2.12632 | 2.96298 | 3.93933 | 5.4207  | 9.04332 | 29.463  | 0.50537 |
| BIA-BIA_BMI                | 1589 | 19.2715 | 4.85536 | 0.04827 | 14.2202 | 15.9187 | 17.9069 | 21.2947 | 29.1104 | 43.3424 | 0.49842 |
| BIA-BIA_BMR                | 1592 | 1188.8  | 280.847 | 813.397 | 907.515 | 1002.96 | 1114.41 | 1301.2  | 1692.4  | 3806.7  | 0.49747 |
| BIA-BIA_DEE                | 1592 | 2000.5  | 615.048 | 1073.45 | 1330.62 | 1600.83 | 1864.86 | 2212.92 | 3130.73 | 7994.08 | 0.49747 |
| BIA-BIA_ECW                | 1592 | 18.9954 | 11.2215 | 1.78945 | 6.86054 | 11.037  | 15.9551 | 24.9752 | 38.9145 | 115.069 | 0.49747 |
| BIA-BIA_FFM                | 1592 | 68.8855 | 29.914  | 28.9004 | 38.9253 | 49.0913 | 60.9631 | 80.8578 | 122.526 | 347.727 | 0.49747 |
| BIA-BIA_FFM1               | 1587 | 14.7588 | 2.1957  | 7.86485 | 12.7842 | 13.4136 | 14.0861 | 15.364  | 18.8683 | 43.7314 | 0.49905 |
| BIA-BIA_FMI                | 1566 | 4.67578 | 3.43821 | 0.10054 | 1.14702 | 2.33693 | 3.71842 | 5.93252 | 11.0777 | 28.2515 | 0.50508 |
| BIA-BIA_Fat                | 1566 | 22.8171 | 20.2481 | 0.38666 | 3.63883 | 8.77408 | 16.1966 | 29.9676 | 65.5142 | 153.82  | 0.50508 |
| BIA-BIA_Frame_num          | 1593 | 1.742   | 0.6813  | 1       | 1       | 1       | 2       | 2       | 3       | 3       | 0.49716 |

Рисунок 3.22 – Статистики за табличними даними (2 частина)

Наступна, третя частина статистичних показників табличних даних наведена на рисунку нижче (рис. 3.23).

|                |      |         |         |         |         |         |         |         |         |         |         |
|----------------|------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| BIA-BIA_ICW    | 1592 | 31.6969 | 11.4608 | 17.555  | 20.5051 | 24.3981 | 28.7014 | 35.1464 | 52.8249 | 138.491 | 0.49747 |
| BIA-BIA_LDM    | 1592 | 18.1932 | 8.13175 | 5.47079 | 9.27148 | 12.846  | 16.2816 | 21.9948 | 32.0028 | 95.6882 | 0.49747 |
| BIA-BIA_LST    | 1592 | 64.4792 | 29.3209 | 23.6201 | 35.2095 | 44.8606 | 56.8376 | 75.8625 | 116.675 | 334.766 | 0.49747 |
| BIA-BIA_SHM    | 1592 | 32.1838 | 18.1718 | 11.3825 | 15.9076 | 21.0347 | 27.271  | 37.9493 | 61.5423 | 223.449 | 0.49747 |
| BIA-BIA_TBW    | 1592 | 50.6923 | 22.2337 | 20.5892 | 28.2804 | 35.7496 | 44.8257 | 59.8235 | 90.8294 | 253.56  | 0.49747 |
| PCIAT-Season   | 2195 | 1.43645 | 1.1345  | 0       | 0       | 0       | 1       | 2       | 3       | 3       | 0.30713 |
| PCIAT-PCIAT_01 | 2192 | 2.37454 | 1.6726  | 0       | 0       | 1       | 2       | 4       | 5       | 5       | 0.30808 |
| PCIAT-PCIAT_02 | 2193 | 2.17465 | 1.70124 | 0       | 0       | 0       | 2       | 4       | 5       | 5       | 0.30777 |
| PCIAT-PCIAT_03 | 2190 | 2.40685 | 1.59495 | 0       | 0       | 1       | 2       | 4       | 5       | 5       | 0.30871 |
| PCIAT-PCIAT_04 | 2191 | 0.83524 | 1.1889  | 0       | 0       | 0       | 0       | 1       | 4       | 5       | 0.3084  |
| PCIAT-PCIAT_05 | 2190 | 2.30365 | 1.70361 | 0       | 0       | 1       | 2       | 4       | 5       | 5       | 0.30871 |
| PCIAT-PCIAT_06 | 2192 | 1.07619 | 1.29004 | 0       | 0       | 0       | 1       | 2       | 4       | 5       | 0.30808 |
| PCIAT-PCIAT_07 | 2190 | 0.58402 | 1.04579 | 0       | 0       | 0       | 0       | 1       | 3       | 5       | 0.30871 |
| PCIAT-PCIAT_08 | 2190 | 1.25297 | 1.34702 | 0       | 0       | 0       | 1       | 2       | 4       | 5       | 0.30871 |
| PCIAT-PCIAT_09 | 2189 | 1.06304 | 1.26392 | 0       | 0       | 0       | 1       | 2       | 4       | 5       | 0.30903 |
| PCIAT-PCIAT_10 | 2192 | 1.31022 | 1.34049 | 0       | 0       | 0       | 1       | 2       | 4       | 5       | 0.30808 |
| PCIAT-PCIAT_11 | 2193 | 1.69403 | 1.54638 | 0       | 0       | 0       | 2       | 3       | 4       | 5       | 0.30777 |

Рисунок 3.23 – Статистики за табличними даними (3 частина)

Наступна, четверта частина статистичних показників табличних даних наведена на рисунку нижче (рис. 3.24).

|                                        |      |         |         |    |         |      |    |    |    |    |         |
|----------------------------------------|------|---------|---------|----|---------|------|----|----|----|----|---------|
| PCIAT-PCIAT_12                         | 2178 | 0.22773 | 0.4501  | 0  | 0       | 0    | 0  | 0  | 1  | 2  | 0.3125  |
| PCIAT-PCIAT_13                         | 2189 | 1.34947 | 1.42639 | 0  | 0       | 0    | 1  | 2  | 4  | 5  | 0.30903 |
| PCIAT-PCIAT_14                         | 2191 | 1.03423 | 1.2946  | 0  | 0       | 0    | 1  | 2  | 4  | 5  | 0.3084  |
| PCIAT-PCIAT_15                         | 2190 | 1.48447 | 1.48926 | 0  | 0       | 0    | 1  | 2  | 4  | 5  | 0.30871 |
| PCIAT-PCIAT_16                         | 2188 | 1.45612 | 1.4994  | 0  | 0       | 0    | 1  | 2  | 5  | 5  | 0.30934 |
| PCIAT-PCIAT_17                         | 2185 | 1.62334 | 1.44923 | 0  | 0       | 0    | 1  | 3  | 4  | 5  | 0.31029 |
| PCIAT-PCIAT_18                         | 2187 | 1.61225 | 1.52521 | 0  | 0       | 0    | 1  | 2  | 5  | 5  | 0.30966 |
| PCIAT-PCIAT_19                         | 2190 | 1.1653  | 1.3491  | 0  | 0       | 0    | 1  | 2  | 4  | 5  | 0.30871 |
| PCIAT-PCIAT_20                         | 2192 | 0.92974 | 1.18279 | 0  | 0       | 0    | 1  | 1  | 4  | 5  | 0.30808 |
| PCIAT-PCIAT_Total                      | 2195 | 27.9171 | 20.3805 | 0  | 0       | 12   | 26 | 41 | 65 | 93 | 0.30713 |
| SDS-Season                             | 2103 | 1.44555 | 1.14162 | 0  | 0       | 0    | 1  | 2  | 3  | 3  | 0.33617 |
| SDS-SDS_Total_Raw                      | 2094 | 41.2163 | 10.3961 | 17 | 28      | 34   | 39 | 46 | 61 | 93 | 0.33902 |
| PreInt_EduHx-Season                    | 2823 | 1.48849 | 1.15358 | 0  | 0       | 0    | 1  | 3  | 3  | 3  | 0.1089  |
| PreInt_EduHx-computerinternet_hoursday | 2633 | 1.06039 | 1.09214 | 0  | 0       | 0    | 1  | 2  | 3  | 3  | 0.16888 |
| sii                                    | 2195 | 0.57995 | 0.77233 | 0  | 0       | 0    | 0  | 1  | 2  | 3  | 0.30713 |
| F6C-F6C_SR                             | 1844 | 8.82893 | 3.31976 | 0  | 3.89375 | 6.75 | 9  | 11 | 14 | 21 | 0.41793 |
| F6C-F6C_SR_Zone                        | 1844 | 0.57809 | 0.494   | 0  | 0       | 0    | 1  | 1  | 1  | 1  | 0.41793 |

Рисунок 3.24 – Статистики за табличними даними (4 частина)

Остання, п'ята частина статистичних показників табличних даних наведена на рисунку нижче (рис. 3.25).

|                                  |      |         |         |      |       |      |      |      |         |      |         |
|----------------------------------|------|---------|---------|------|-------|------|------|------|---------|------|---------|
| Fitness_Endurance_Total_Time_Sec | 3165 | 87.7343 | 197.761 | 0    | 0     | 0    | 0    | 0    | 583     | 987  | 0.00095 |
| PAQ_Total                        | 1754 | 2.5033  | 0.81435 | 0.58 | 1.233 | 1.88 | 2.46 | 3.09 | 3.91385 | 4.79 | 0.44634 |
| PAQ_Season                       | 1754 | 1.46978 | 1.17759 | 0    | 0     | 0    | 1    | 3    | 3       | 3    | 0.44634 |
| PAQ_Zone                         | 1754 | 0.38483 | 0.48669 | 0    | 0     | 0    | 0    | 1    | 1       | 1    | 0.44634 |

Рисунок 3.25 – Статистики за табличними даними (5 частина)

Також, важливою є візуалізація даних, для більш точного розуміння повної картини.

За віковою категорією більшість досліджуваних – це діти та підлітки від восьми до чотирнадцяти років, що доводить факт формування Інтернет-залежності в ранньому віці [7; 8] (рис. 3.26).

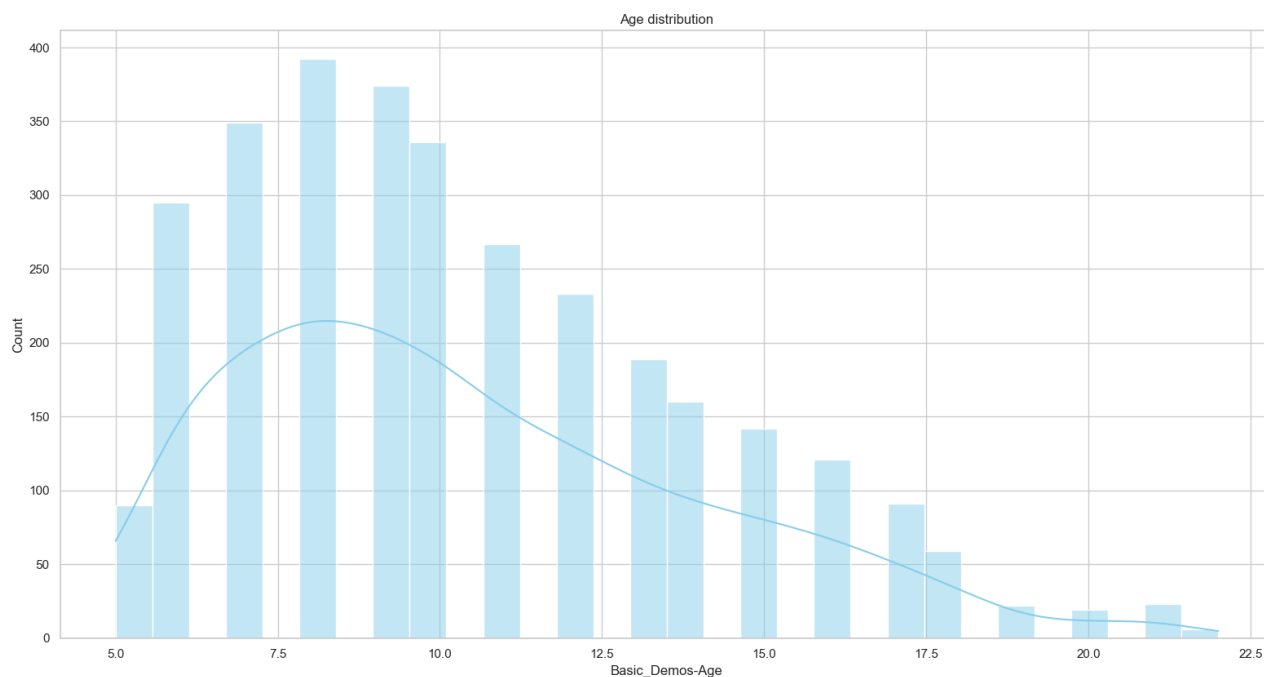


Рисунок 3.26 – Гістограма «Вік»

Розподіл значень індексу маси тіла близький до норми, за винятком деяких екземплярів з відхиленнями в обидва боки (рис. 3.27).

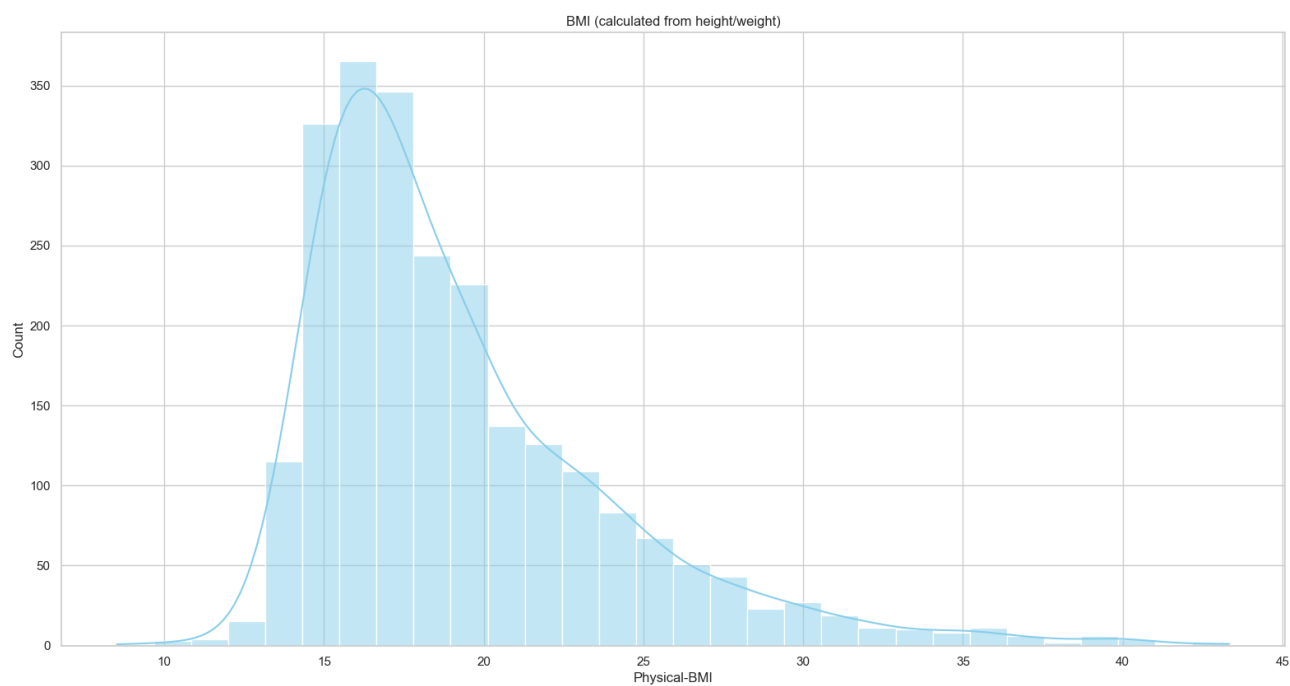


Рисунок 3.27 – Гістограма «Індекс маси тіла»

Розподіл значень ознак систолічного та діастолічного тиску також близький до норми, за винятком деяких екземплярів з відхиленнями в обидва боки (рис. 3.28 – рис. 3.29).

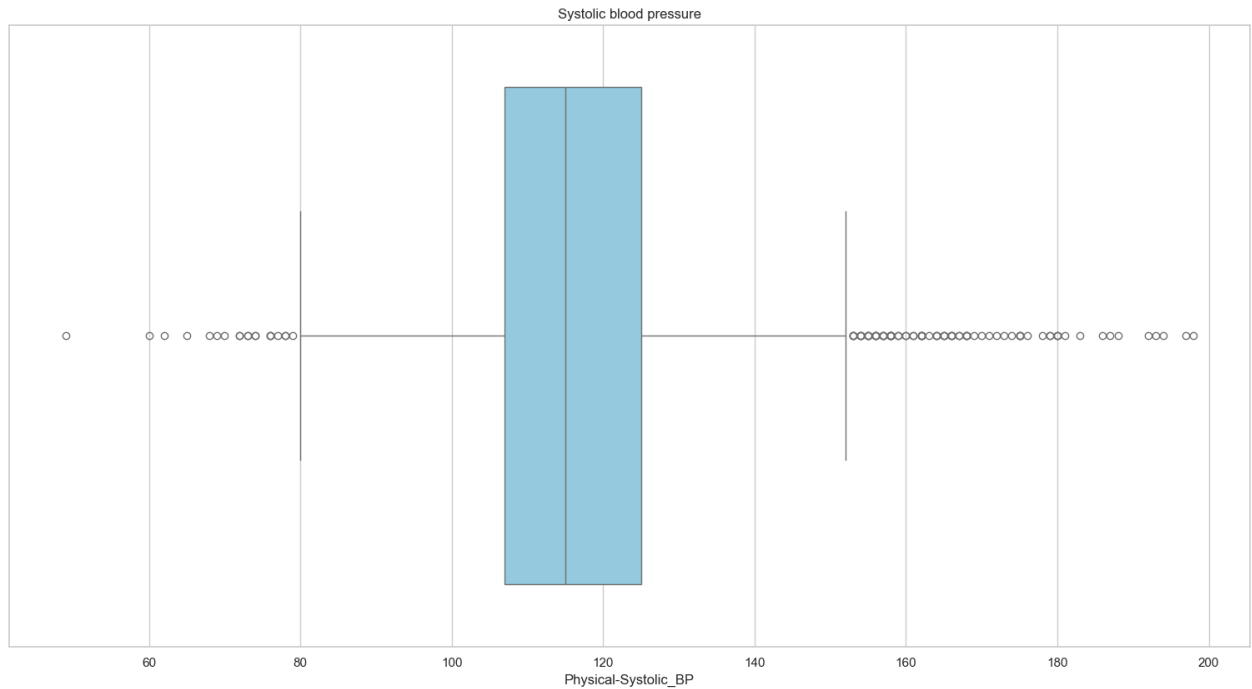


Рисунок 3.28 – Боксплот «Систолічний тиск»

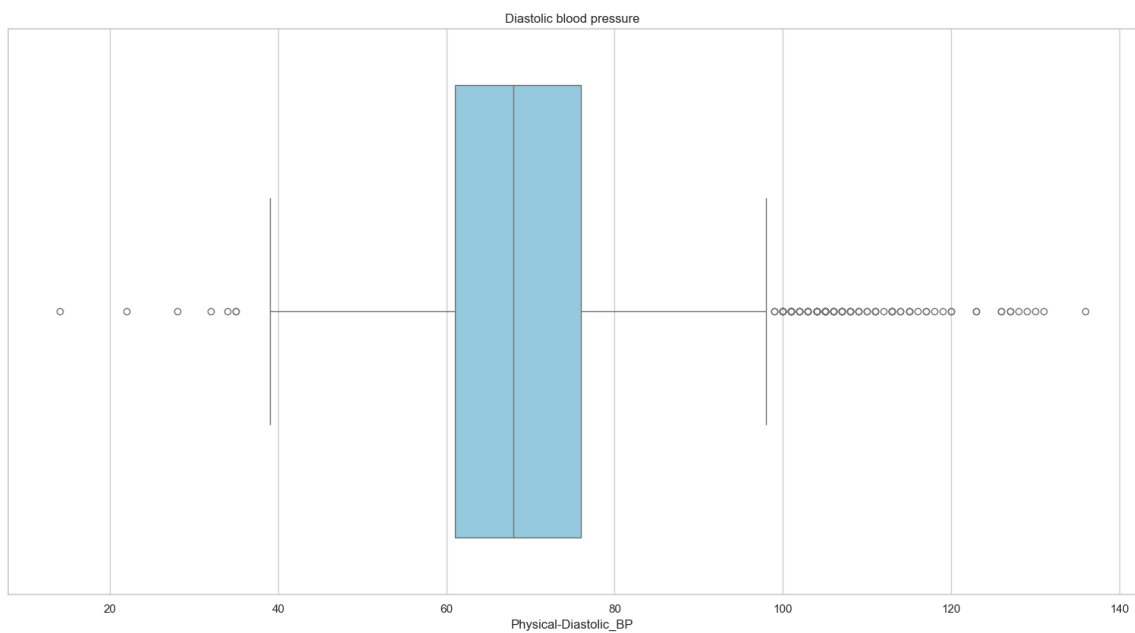


Рисунок 3.29 – Боксплот «Діастолічний тиск»

Розподіл значень частоти серцевих скорочень є, аналогічно до попередніх ознак, дуже близьким до норми (рис. 3.30).

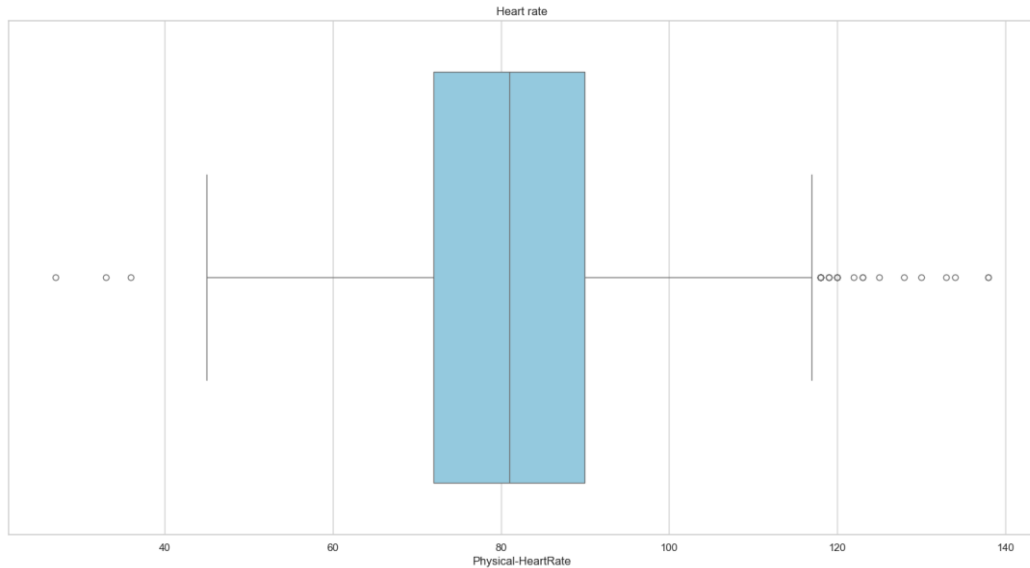


Рисунок 3.30 – «Частота серцевих скорочень»

Переходячи до результатів психологічних вимірів – значення шкали CGAS прямують до середніх або ж високих, хоча є і винятки, які можуть бути індикаторами погіршення рівня адаптивних можливостей дитини (рис. 3.31).

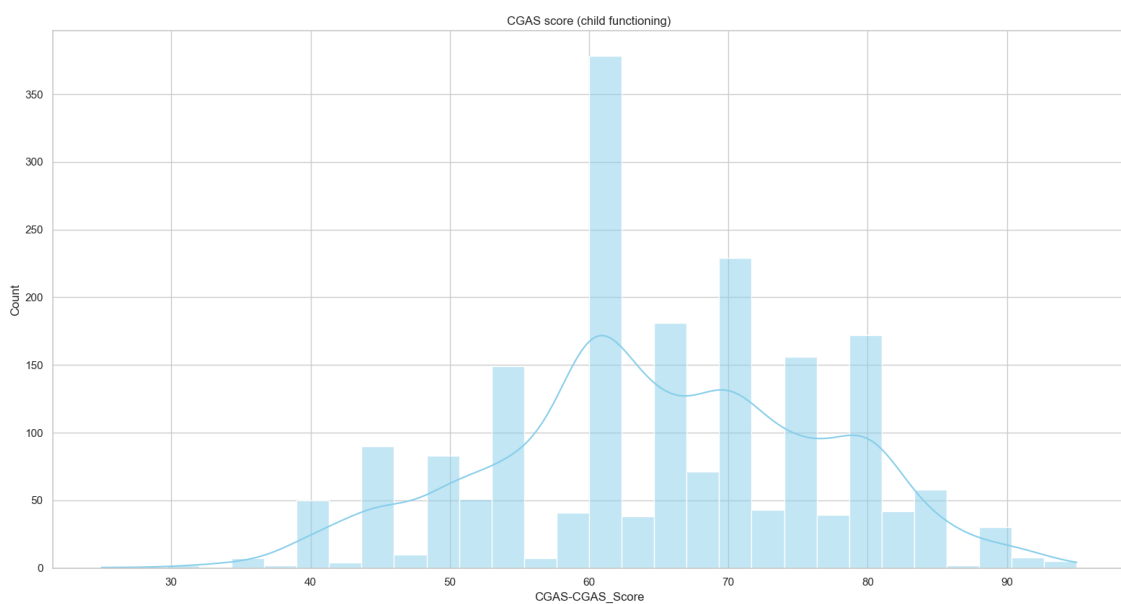


Рисунок 3.31 – Гістограма «Глобальна оцінка функціонування CGAS»

Результати PCIAT-опитувальника мають зміщення від нульових значень певної групи учасників, що може бути прямим вказівником на наявність залежності від Інтернету, хоча й найбільший пік графіку припадає на нуль (рис. 3.32).

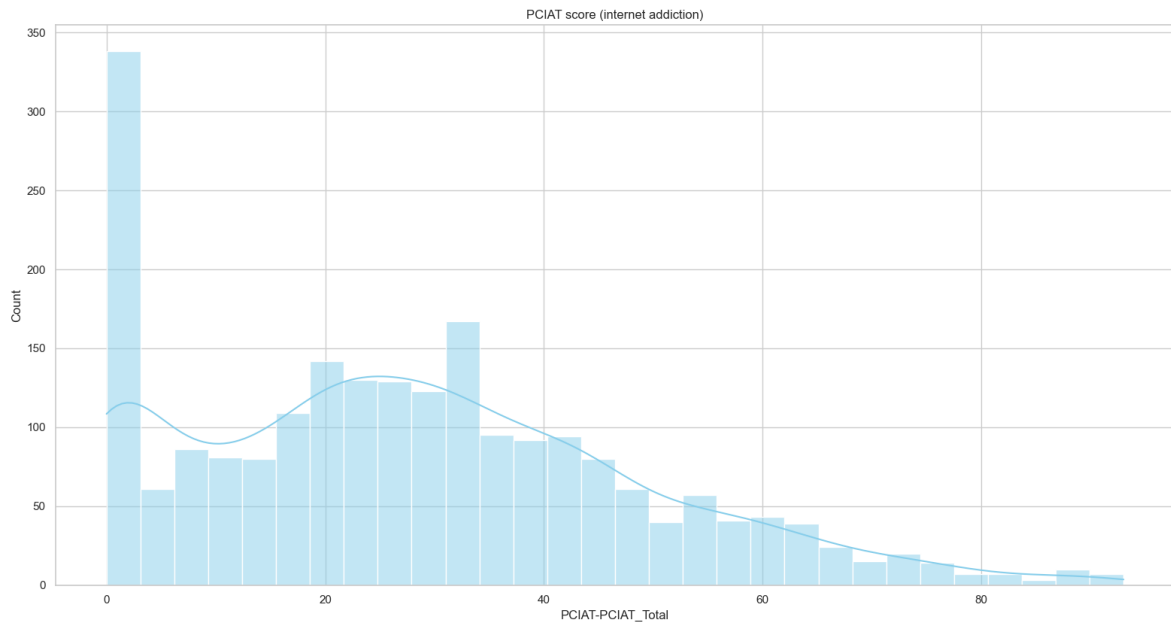


Рисунок 3.32 – Гістограма «Індекс інтернет-залежності PCIAT»

За значеннями SDS можна визначити групу учасників з певними симптомами тривожно-депресивного стану, а PAQ-результати вказують на значний розкид за рівнем фізичної активності (рис. 3.33).

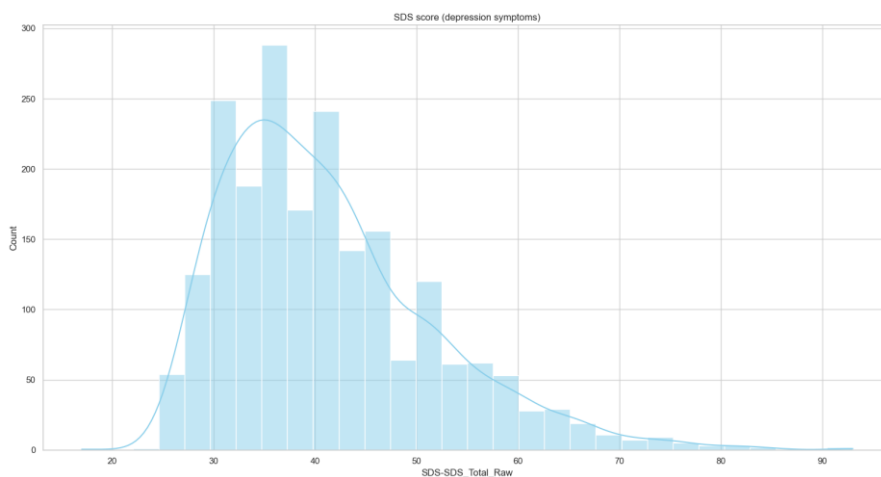


Рисунок 3.33 – Гістограма «Депресивні симптоми SDS»

Тести фізичної витривалості дають змогу зрозуміти, що піддослідні не мають високого рівня фізичної підготовки, і лише незначна група показує прийнятні результати (рис. 3.34 – рис. 3.35).

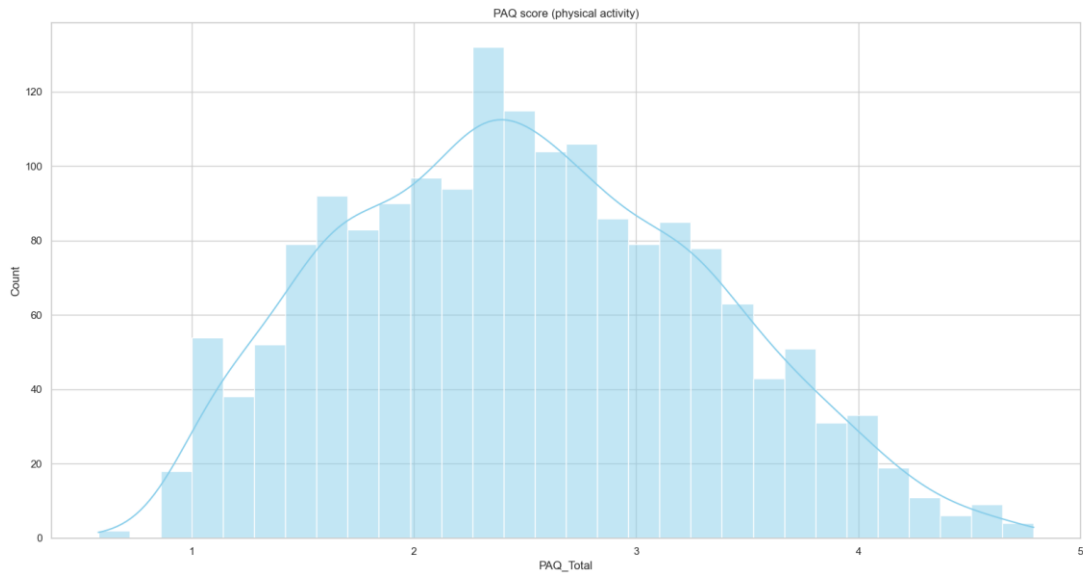


Рисунок 3.34 – Гістограма «Рівень фізичної активності PAQ»

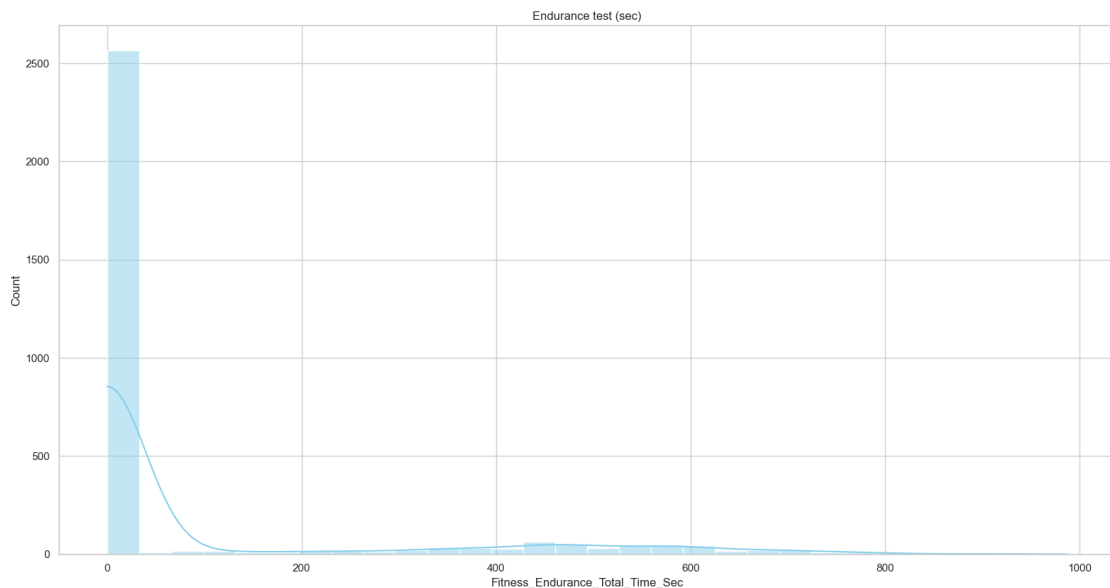


Рисунок 3.35 – Гістограма «Тест на витривалість»

З іншого боку тест на гнучкість «Sit and Reach» розподіляється рівномірніше по всьому графіку (рис. 3.36).

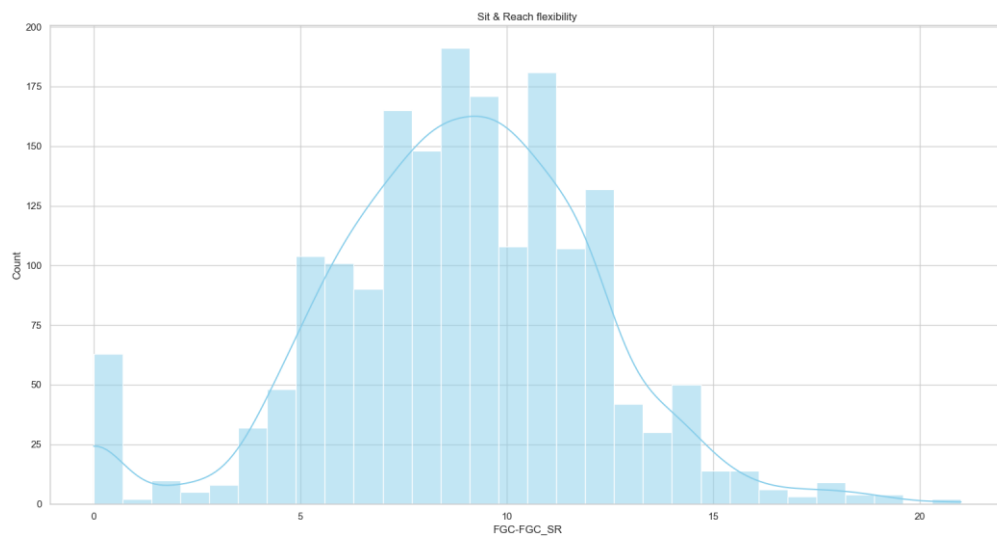


Рисунок 3.36 – Гістограма «Тест на гнучкість Sit&Reach»

Останніми найбільш інформативними даними є ознаки, пов'язані зі складом тіла дітей, відображаючи як нормальні значення у жировій та м'язовій частинах тіла, так і надлишкові (рис. 3.37).

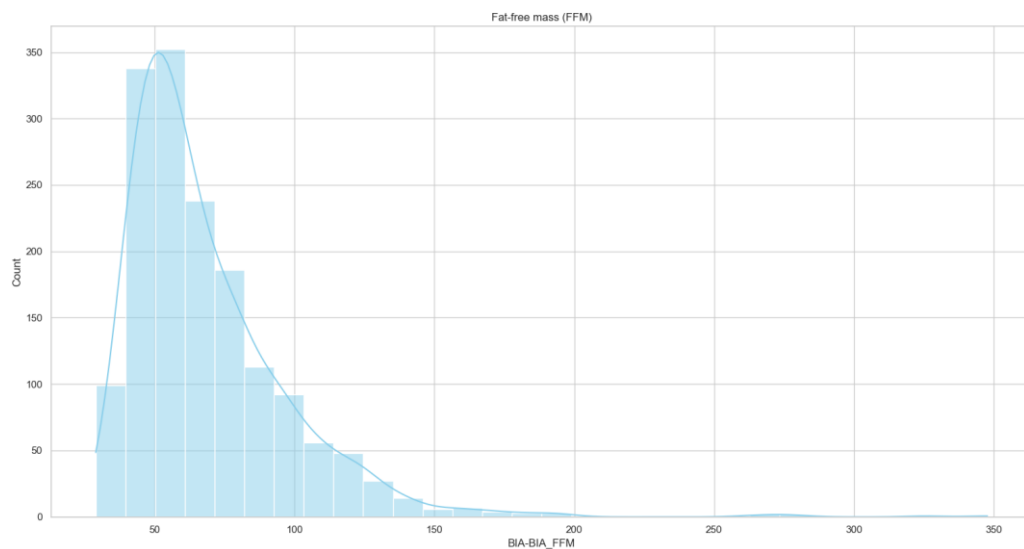


Рисунок 3.37 – Гістограма «BIA-показники складу тіла»

Підсумовуючи, маємо такі результати: більшість учасників – діти віком 8-14 років, у яких є певні ознаки Інтернет-залежності та деякі симптоми тривожно-депресивного стану. Фізіологічні показники в більшості випадків

мають нормальні значення, фізична підготовка загалом низька, склад тіла здебільшого в межах норми.

#### **3.2.4. Оброблення пропущених значень**

На попередньому етапі роботи, в декількох місцях було зазначено, що значна частина даних містила в собі пропущені значення. Це явище може відбуватися за багатьма різними причинами, такими як різноманітні технічні збої, недосконале або неповне заповнення даних тощо. Ігнорування пропущених значень може призвести до неточності отримуваних результатів, а банальне видалення рядків даних повністю – до критичного зменшення вибірки. Саме тому важливою частиною обробки даних являється заповнення пропусків, використовуючи вже наявну інформацію.

Таким методами є як звичайна заміна пропущених значень на середнє чи медіанне по всіх наявних даних, так і більш глибокі, сучасні алгоритми, які враховують взаємозв'язки змінних між собою.

Найбільш пріоритетною задачею є заповнення пропущених значень ключової змінної опитувальника PCİAT (за значеннями якого в подальшому по встановлених межах можна визначити клас SII). Саме тому, варто спробувати різні варіанти алгоритмів та їх параметрів, для знаходження найоптимальнішого з них, в рамках поставленої задачі.

Таким чином, було створено програмний код, який створював еталонну вибірку екземплярів, які б не містили пропусків по жодному з двадцяти пунктів опитувальника PCİAT. Далі в ньому створювалися десятивідсоткові пропуски випадковим чином, що давало змогу отримати уявлення про якість відновлюваних даних різними методами. Результати оцінювались за середньою абсолютною помилкою та коренем середньоквадратичної помилки.

Було випробувано два методи заповнення пропущених значень, а саме імпутер за алгоритмом KNN з різною кількістю сусідів та множинна імпутація за допомогою ланцюгових рівнянь MICE з різними базовими алгоритмами [39].

Результати були окремо винесені в таблицю (рис. 3.38) та візуалізовано на графіку (рис. 3.39).

| Method               | MAE    | RMSE   |
|----------------------|--------|--------|
| MICE (RandomForest)  | 0.5929 | 0.8836 |
| MICE (BayesianRidge) | 0.5997 | 0.8652 |
| MICE (ExtraTrees)    | 0.6066 | 0.9067 |
| KNN (k=10)           | 0.6174 | 0.9171 |
| KNN (k=7)            | 0.6227 | 0.9334 |
| KNN (k=5)            | 0.6322 | 0.9562 |
| KNN (k=3)            | 0.6532 | 0.9971 |

Рисунок 3.38 – Результати порівняння методів заповнення пропущених значень

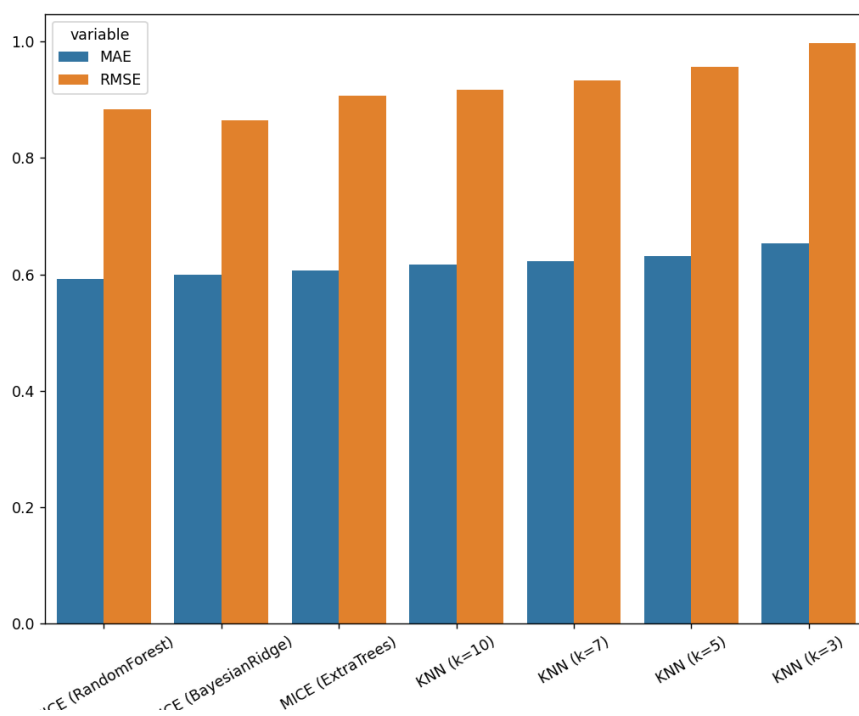


Рисунок 3.39 – Графік порівняння методів для заповнення пропущених значень

Відновлення синтетично пропущених значень вказує на стабільно кращі результати при використанні методу MICE з різними базовими алгоритмами ніж при використанні KNN, з меншою похибкою відновлення.

Найкращим з усіх виявився метод MICE, в основі якого лежить алгоритм випадкового лісу, який і буде використовуватися для заповнення пропусків у відповідях до тесту PCIAT.

Для всіх учасників дослідження (за винятком тих, у кого абсолютно всі з двадцяти запитань були пропущені – їх було виключено) дані відповідей тесту PCIAT були заповнені, кількість пропусків стала рівна нулю.

Таким чином далі було перерозраховано для кожного з піддослідних значення змінної «PCIAT-PCIAT\_Total», як суму з двадцяти значень, а також виведено за завчасно встановленими межами значення індексу SII.

Усі ж інші дані, в подальшому, при роботі з моделями машинного навчання, будуть заповнюватися методом множинної імпультації, що відновлює пропущені дані через послідовну побудову регресійних рівнянь між змінними. Цей метод поєднує точність та практичну ефективність, враховуючи значний розмір датасету, забезпечуючи повноту спостережень та зберігаючи стабільність результатів.

### **3.2.5. Видалення надлишкових ознак**

Наступним кроком варто позбутися надлишкових та нерелевантних ознак. Треба позбавитись з набору даних ідентифікаторів, ознак, що не мають науково-теоретичного обґрунтування, ознак із завищеною кореляцією, та тих, що не мають інформаційної цінності для подальшого дослідження. Тим самим кількість ознак ще скоротиться, полегшуючи подальшу роботу та точність і якість фінальних результатів, оскільки велика кількість змінних, навіть неважливих, може досить сильно їх спотворювати.

Таким чином, було видалено наступні ознаки:

1. Змінні «id» та «has\_actigraphy» – ідентифікатор учасника експерименту та флаг наявності актиграфічних даних.

Рисунок 3.40 – Матриця кореляцій між змінними

Варто визначити всі кореляції між змінними, які є більшими за 0.95 (завищені кореляції), а також зробити висновок, які з ознак можуть походити одна від одної та є кандидатами на видалення.

Таким чином, перша частина попарних кореляцій між змінними на рисунку нижче (рис. 3.41).

| Variable 1                  | Variable 2                       | Correlation | Suggested_to_drop           |
|-----------------------------|----------------------------------|-------------|-----------------------------|
| BIA-BIA_BMR                 | BIA-BIA_FFM                      | 1           | Either                      |
| BIA-BIA_BMR                 | BIA-BIA_LST                      | 0.997       | Either                      |
| BIA-BIA_FFM                 | BIA-BIA_LST                      | 0.997       | Either                      |
| BIA-BIA_BMR                 | BIA-BIA_TBW                      | 0.995       | Either                      |
| BIA-BIA_FFM                 | BIA-BIA_TBW                      | 0.995       | Either                      |
| BIA-BIA_LST                 | BIA-BIA_TBW                      | 0.993       | Either                      |
| BIA-BIA_ICW                 | BIA-BIA_TBW                      | 0.98        | Either                      |
| BIA-BIA_ECW                 | BIA-BIA_TBW                      | 0.979       | Either                      |
| Fitness_Endurance-Max_Stage | Fitness_Endurance_Total_Time_Sec | 0.978       | Fitness_Endurance-Max_Stage |
| BIA-BIA_BMR                 | BIA-BIA_ICW                      | 0.976       | Either                      |

Рисунок 3.41 – Визначення завищених кореляцій між змінними (частина 1)

Наступна, друга частина попарних кореляцій між змінними на рисунку нижче (рис. 3.42).

|                   |                 |       |                 |
|-------------------|-----------------|-------|-----------------|
| BIA-BIA_FFM       | BIA-BIA_ICW     | 0.976 | Either          |
| Physical-BMI      | BIA-BIA_BMI     | 0.975 | BIA-BIA_BMI     |
| BIA-BIA_BMR       | BIA-BIA_ECW     | 0.973 | Either          |
| BIA-BIA_ECW       | BIA-BIA_FFM     | 0.973 | Either          |
| BIA-BIA_ICW       | BIA-BIA_LST     | 0.973 | Either          |
| BIA-BIA_ECW       | BIA-BIA_LST     | 0.972 | Either          |
| minutes_mvpa_mean | mvpa_share_mean | 0.969 | mvpa_share_mean |
| BIA-BIA_ICW       | BIA-BIA_SMM     | 0.961 | Either          |
| BIA-BIA_FFM       | BIA-BIA_LDM     | 0.961 | Either          |
| BIA-BIA_BMR       | BIA-BIA_LDM     | 0.961 | Either          |
| BIA-BIA_BMR       | BIA-BIA_SMM     | 0.96  | Either          |

Рисунок 3.42 – Визначення завищених кореляцій між змінними (частина 2)

Остання, третя частина попарних кореляцій між змінними на рисунку нижче (рис. 3.43).

|                       |                       |       |                       |
|-----------------------|-----------------------|-------|-----------------------|
| BIA-BIA_FFM           | BIA-BIA_SMM           | 0.96  | Either                |
| rhythm_amplitude_mean | rhythm_stability_mean | 0.959 | rhythm_amplitude_mean |
| BIA-BIA_BMI           | BIA-BIA_FMI           | 0.957 | BIA-BIA_FMI           |
| BIA-BIA_SMM           | BIA-BIA_TBW           | 0.957 | Either                |
| BIA-BIA_LST           | BIA-BIA_SMM           | 0.956 | Either                |
| BIA-BIA_LDM           | BIA-BIA_LST           | 0.954 | Either                |

Рисунок 3.43 – Визначення завищених кореляцій між змінними (частина 3)

В результаті отримуємо:

1. Попарні кореляції між ознаками «BIA-» вказують на те, що вони є похідними, тобто розраховуються одна через одну, утворюючи спільний кластер. Залишаємо лише змінну «BIA-BIA\_LST», як найбільш інтерпретовану змінну.

2. Ознаки «BIA-BIA\_BMI» і «BIA-BIA\_FMI» мають сильний зв'язок із ознакою «Physical-BMI», тому лишаємо «Physical-BMI», як більш інтерпретовану та краще вимірювану змінну.

3. Результати фітнес-тестів «Fitness\_Endurance\_Total\_Time\_Sec» та «Fitness\_Endurance-Max\_Stage» – мають високу попарну кореляцію. Залишаємо більш універсальний час у секундах.

4. Ознаки рухової активності «minutes\_mvpa\_mean» та «mvpa\_share\_mean» – аналогічно. Залишаємо більш зрозумілий показник у хвилинах.

5. Ознаки ритмів «rhythm\_amplitude\_mean» та «rhythm\_stability\_mean» – аналогічно. Залишаємо більш узагальнену ознаку стабільності.

### 3.2.6. Визначення важливості ознак

Останнім етапом проведемо визначення важливості ознак, який може знадобитися в наступних кроках роботи, зокрема при моделюванні. Маючи ключову ознаку «PCIAT-PCIAT\_Total» (та ознаку «SII», яка на пряму будується з неї), визначимо найсильніші кореляції всіх інших змінних із нею (рис. 3.44).

| Variable                               | Correlation |
|----------------------------------------|-------------|
| Physical-Height                        | 0.418       |
| Basic_Demos-Age                        | 0.415       |
| PreInt_EduHx-computerinternet_hoursday | 0.381       |
| Physical-Weight                        | 0.365       |
| BIA-BIA_LST                            | 0.329       |
| Physical-Waist_Circumference           | 0.318       |
| BIA-BIA_DEE                            | 0.313       |
| FGC-FGC_CU                             | 0.308       |
| BIA-BIA_Fat                            | 0.279       |
| enmo_mean_mean                         | 0.254       |

Рисунок 3.44 – Значення найбільш корельованих ознак з «PCIAT-PCIAT\_Total»

Проаналізувавши різні модифікації змінних (приведення до квадрату, знаходження квадратного кореня та логарифму тощо), було визначено, що вони не містять вищих значень кореляції з ключовою змінною, ніж початкові, «сирі» ознаки.

### 3.3. Обґрунтування вибору фінальної моделі

Вибір моделі, для її подальшого використання в програмному продукті – без перебільшення – найважливіший крок усього дослідження. Саме тому, цей етап вимагає найбільшої уваги та ресурсів під час роботи.

### 3.3.1. Загальна схема роботи

Як вже відомо з розділу щодо засобів реалізації, тестуватися будуть деякі види різноманітних алгоритмів та методів. Таким чином, для повноцінного розуміння якості їх роботи, всі вони мають бути в однакових умовах – підлягати єдиній схемі дій. Як класифікатори, так і регресори з нейронними мережами будуть проходити спільний порядок кроків.

1. Завантаження очищеного та обробленого набору даних, із заповненими пропусками в ключових ознаках. З таблиці прибираються ключові ознаки та ознаки, від яких ключові залежать напряду.

2. Заповнення пропущених даних в неключових ознаках. Використовується метод множинної імпутації Iterative Imputer [40].

3. Масштабування числових ознак для чутливих до різниці у діапазонах алгоритмів, стандартизуючи дані за використанням інструменту StandardScaler, що гарантує приведення змінних до дисперсії, рівній одиниці та середньому, рівному нулю [41].

4.1. Балансування класів у задачі класифікації. Через дуже нерівномірний розподіл спостережень між класами, зокрема в класі «0» (найбільша кількість) та в класі «3» (майже немає) використовується метод Synthetic Minority Oversampling Technique (SMOTE), який за заданою пропорцією синтетично додає до вибірки нові спостереження до класу «3», за використанням їх найближчих сусідів.

4.2. Балансування класів у задачі регресії не проводиться, оскільки ключова змінна «PCIAT-PCIAT\_Total» не являється категоріальною. В подальшому (для порівняння результатів) переведення даної ознаки в класову «SII» виконуватиметься за завчасно вказаними межами [0; 30; 49; 79; 100], кожна з яких відповідає ступеню проблематичності використання Інтернету дитиною [37].

5. Оцінювання алгоритмів за використанням п'ятифолдної стратифікованої крос-валідації, в якій весь набір даних поділяється на п'ять

рівних частин, не втрачаючи балансу між класами. На перших чотирьох проходах відбувається навчання моделей, а на п'ятому – тестування.

6. Розрахунок метрики якості, а саме коефіцієнту згоди Каппа, з використанням квадратичних ваг. Таким рішенням ми враховуємо відстань між неякісно отриманими результатами (класами), що є суттєвим для поставленої задачі. Фінальні оцінки зберігаються, після чого розраховується середнє значення по всіх ітераціях, що дає повноцінну картину якості використання кожного з методів та алгоритмів.

### **3.3.2. Класифікаційні моделі**

Першими до розгляду було вирішено взяти моделі класифікації. Для отримання початкового розуміння як кожен з алгоритмів буде працювати на нашому датасеті в рамках поставленої задачі тестуються початкові версії.

Для того щоб проаналізувати широкий спектр методів машинного навчання для роботи обрані як звичайні, лінійні підходи до класифікації, так і складніші:

1. Метод логістичної регресії (Logistic Regression);
2. Метод дерева рішень для класифікації (Decision Tree Classifier);
3. Метод k-найближчих сусідів для класифікації (K-Nearest Neighbors (KNN) Classifier);
4. Метод опорних векторів для класифікації (Support Vector Classifier (SVC));
5. Метод випадкового лісу для класифікації (Random Forest Classifier);
6. Метод градієнтного бустингу для класифікації (Gradient Boosting Classifier);
7. Метод XGBoost для класифікації (XGBoost Classifier);
8. Метод LightGBM для класифікації (LightGBM Classifier);
9. Метод CatBoost для класифікації (CatBoost Classifier).

З результатами роботи алгоритмів без підбору параметрів (відсортовані за зниженням точності) «з коробки» можна ознайомитися нижче (рис. 3.45):

| Classifier                   | Kappa (Predict SII) |
|------------------------------|---------------------|
| Logistic Regression (Clf)    | 0.379               |
| Gradient Boosting (Clf)      | 0.372               |
| CatBoost (Clf)               | 0.367               |
| Support Vector Machine (Clf) | 0.334               |
| XGBoost (Clf)                | 0.331               |
| LightGBM (Clf)               | 0.328               |
| Random Forest (Clf)          | 0.302               |
| K-Nearest Neighbors (Clf)    | 0.253               |
| Decision Tree (Clf)          | 0.219               |

Рисунок 3.45 – Результати роботи базових класифікаційних моделей

Аналізуючи отримані результати, можна дійти висновку, що найкращі результати за значеннями коефіцієнтів згоди Каппа показали алгоритм логістичної регресії (0.379), звичайний метод градієнтного бустингу (0.372) та ансамблевий градієнтний класифікатор CatBoost (0.367). Інші алгоритми видали посередні результати, а найгіршими виявились метод найближчих сусідів (0.253) та дерево рішень (0.219). Можна зробити висновок, що всі базові підходи видали на виході довірчий рівень узгодженості отриманих результатів з реальними.

Подальша робота передбачає покращення отриманих результатів базових моделей через підбір оптимального набору параметрів для кожного з алгоритмів. Таким чином, після цього отримуємо наступні результати (рис. 3.46):

| Classifier                   | Kappa (Predict SII) |
|------------------------------|---------------------|
| Random Forest (Clf)          | 0.399               |
| Logistic Regression (Clf)    | 0.383               |
| CatBoost (Clf)               | 0.371               |
| Gradient Boosting (Clf)      | 0.364               |
| Support Vector Machine (Clf) | 0.363               |
| XGBoost (Clf)                | 0.348               |
| LightGBM (Clf)               | 0.335               |
| Decision Tree (Clf)          | 0.292               |
| K-Nearest Neighbors (Clf)    | 0.272               |

Рисунок 3.46 – Результати роботи класифікаційних моделей з підібраними параметрами

Порівнюючи результати, отримані внаслідок підбору параметрів, з базовими конфігураціями моделей, можна прийти висновку, що найбільше покращення здобув алгоритм випадкового лісу (з 0.302 до 0.399), майже досягнувши наступного, помірного рівня узгодженості. Залишившись одним з найкращих методів, невеликий приріст значення коефіцієнту Каппа (з 0.379 до 0.383) показала класична логістична регресія. Градієнтні ансамблеві методи досить непогано покращили свої результати, тим самим наздоганяючи фаворитів. Натомість, результати алгоритму градієнтного бустингу, за будь-якої зміни гіперпараметрів тільки погіршувалися. Найгірші результати, як і на минулому етапі показали підхід дерева рішень (0.292) та найближчих сусідів (0.272).

Останнім етапом роботи з класифікаційними моделями є поєднання декількох з них в ансамблі, використовуючи різноманітні комбінації, з метою покращення попередніх результатів. Використовуватимемо два підходи – стекінгу та голосування. З результатами даного етапу роботи можна ознайомитися нижче (рис. 3.47):

| Classifier                                                                                                                                          | Kappa (Predict SII) |
|-----------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| Stacking (Random Forest (Clf) + XGBoost (Clf) + Gradient Boosting (Clf) + CatBoost (Clf) + LightGBM (Clf) + SVM (Clf) -> Logistic Regression (Clf)) | 0.409               |
| Stacking (Random Forest (Clf) + Gradient Boosting (Clf) + Logistic Regression (Clf))                                                                | 0.396               |
| Voting (Logistic Regression (Clf) + Decision Tree (Clf) + K-Nearest Neighbors (Clf))                                                                | 0.350               |
| Voting (Random Forest (Clf) + Gradient Boosting (Clf) + LightGBM (Clf))                                                                             | 0.347               |

Рисунок 3.47 – Результати роботи ансамблів з класифікаційних моделей з підібраними параметрами

За результатами останнього етапу роботи з класифікаційними моделями, можна побачити, що найкраще з поставленою задачею прогнозування впорався ансамбль стекінгу, що включає в собі такі алгоритми, як випадкового лісу, градієнтного бустингу, класифікаторів машини опорних векторів, CatBoost, XGBoost та LightGBM з логістичною регресією в якості мета-класифікатора (0.409), який першим з усіх досліджуваних алгоритмів досяг помірного рівня узгодженості отриманих результатів із реальними.

Серед ансамблів на основі голосування результати були нижчими, що можна пояснити певними факторами, такими як низькі, відносно інших індивідуальні показники моделей, гірша комбінація алгоритмів тощо.

Саме тому, можна стверджувати, що в рамках поставленої класифікаційної задачі прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників, стекінгові ансамблі мають досить суттєву перевагу як над голосувальними ансамблями, так і над поодинокими моделями класифікації.

### 3.3.3. Регресійні моделі

Наступними до аналізу було взято моделі регресії. Основною метою даного етапу є оцінка якості прогнозування різними алгоритмами сумарного індексу PCIAТ опитувальника та подальшим переведенням цих результатів у значення ознаки класифікації рівня Інтернет-залежності.

Як і у випадку з класифікацією, були обрані класичні поодинокі та складніші ансамблеві підходи, список яких можна побачити нижче:

#### 1. Метод лінійної регресії (Linear Regression);

2. Метод дерева рішень для регресії (Decision Tree Regressor);
3. Метод k-найближчих сусідів для регресії (K-Nearest Neighbors (KNN) Regressor);
4. Метод опорних векторів для регресії (Support Vector Regressor (SVR));
5. Метод випадкового лісу для регресії (Random Forest Regressor);
6. Метод градієнтного бустингу для регресії (Gradient Boosting Regressor);
7. Метод XGBoost для регресії (XGBoost Regressor);
8. Метод LightGBM для регресії (LightGBM Regressor);
9. Метод CatBoost для регресії (CatBoost Regressor).

З результатами роботи алгоритмів без підбору параметрів (відсортовані за зниженням точності) «з коробки» можна ознайомитися нижче (рис. 3.48):

| Regressor                    | Kappa (Bins from PCIAT) |
|------------------------------|-------------------------|
| CatBoost (Reg)               | 0.396                   |
| Gradient Boosting (Reg)      | 0.394                   |
| Linear Regression (Reg)      | 0.381                   |
| Random Forest (Reg)          | 0.379                   |
| LightGBM (Reg)               | 0.371                   |
| XGBoost (Reg)                | 0.355                   |
| Support Vector Machine (Reg) | 0.326                   |
| K-Nearest Neighbors (Reg)    | 0.288                   |
| Decision Tree (Reg)          | 0.231                   |

Рисунок 3.48 – Результати роботи базових регресійних моделей

Аналізуючи отримані результати, можна дійти висновку, що ситуація з регресійними моделями до підбору оптимальних параметрів є кращою ніж в класифікаційних. Найкращі результати за значеннями коефіцієнтів узгодження Каппа показали регресор CatBoost (0.396) та базовий метод градієнтного бустингу (0.394), які майже одразу досягли помірного рівня узгодженості результатів. Усі інші методи, за винятком машини опорних

векторів (0.326), методів k-найближчих сусідів (0.288) та дерева рішень (0.231) (аналогічно з класифікацією) також показали конкурентні результати.

Подальша робота, як і у попередньому випадку, полягає в покращенні отриманих результатів базових моделей через підбір оптимального набору гіперпараметрів для кожного з алгоритмів. Таким чином, після цього отримуємо наступні результати (рис. 3.49):

| Regressor                    | Kappa (Bins from PCIAT) |
|------------------------------|-------------------------|
| CatBoost (Reg)               | 0.417                   |
| Random Forest (Reg)          | 0.394                   |
| LightGBM (Reg)               | 0.383                   |
| Linear Regression (Reg)      | 0.381                   |
| Gradient Boosting (Reg)      | 0.367                   |
| XGBoost (Reg)                | 0.361                   |
| Support Vector Machine (Reg) | 0.356                   |
| K-Nearest Neighbors (Reg)    | 0.346                   |
| Decision Tree (Reg)          | 0.316                   |

Рисунок 3.49 – Результати роботи регресійних моделей з підібраними параметрами

Найкращою моделлю за результатами всіх етапів став регресор CatBoost (0.417) з помірним рівнем узгодженості результатів з реальними даними. Більша частина алгоритмів, як і на етапі тестування класифікаційних моделей з підібраними гіперпараметрами, показала більш точні результати за свої базові версії, хоча значення коефіцієнту згоди Каппа лінійної регресії (0.381) покращити не вдалось, а метод градієнтного бустингу за будь-яких змін тільки погіршував свої показники (з 0.394 до 0.367).

Останнім етапом роботи з регресійними моделями є поєднання декількох з них в ансамблі, використовуючи різноманітні комбінації, з метою покращення попередніх результатів. Використовуватимемо два підходи –

стекінгу та голосування. З результатами даного етапу роботи можна ознайомитися нижче (рис. 3.50):

| Regressor                                                                                                   | Kappa (Bins from PCIAT) |
|-------------------------------------------------------------------------------------------------------------|-------------------------|
| Voting (CatBoost (Reg) + Random Forest (Reg) + Linear Regression (Reg))                                     | 0.398                   |
| Stacking (CatBoost (Reg) + LightGBM (Reg) + XGBoost (Reg) -> Linear Regression (Reg))                       | 0.398                   |
| Voting (CatBoost (Reg) + LightGBM (Reg) + XGBoost(Reg))                                                     | 0.395                   |
| Stacking (Random Forest (Reg) + CatBoost (Reg) + LightGBM (Reg) + XGBoost (Reg) -> Linear Regression (Reg)) | 0.386                   |

Рисунок 3.50 – Результати роботи ансамблів з регресійних моделей з підібраними параметрами

Згідно з результатами даного, завершального етапу роботи над задачею регресії, можна побачити, що на відміну від класифікаційних ансамблів стекінгу та регресійного методу CatBoost з оптимізованими гіперпараметрами, вищезазначеним ансамблям не вдалося досягнути планки помірної узгодженості отриманих результатів із реальними даними.

Хоча ансамблі з декількох моделей і показали конкурентні результати, на етапі дослідження задачі регресії найкраще впорався з прогнозуванням рівня проблемного використання Інтернету дітьми регресор CatBoost з підібраними гіперпараметрами моделі.

### 3.3.4. Нейронні мережі

На заключному етапі практичних експериментів над методами та алгоритмами машинного навчання є застосування штучно створених нейронних мереж для поставленої задачі прогнозування. Це може бути важливим кроком дослідження, оскільки нейромережі в певних випадках, здатні досить точно відтворювати складні нелінійні взаємозв'язки між змінними, що може стати корисною особливістю для досліджуваного датасету з великою кількістю різноманітних даних.

Першопочатково було створено дві базові архітектури – для задачі класифікації та задачі регресії (рис. 3.51).

| Початкова нейромережа |                                                                 |
|-----------------------|-----------------------------------------------------------------|
| Рівень                | Конфігурація                                                    |
| Вхід                  | N                                                               |
| Hidden                | Linear ( $N \rightarrow 128$ ) $\rightarrow$ ReLU               |
| Вихід                 | Linear ( $128 \rightarrow 1$ ) / Linear ( $128 \rightarrow 4$ ) |

Рисунок 3.51 – Архітектура початкової нейронної мережі

Прогнозування проводилось аналогічно до моделей машинного навчання класифікації та регресії, розглянутих на попередніх етапах. В обох випадках моделі мають один прихований шар з функцією активації ReLU, використовується стандартизація ознак та стратегія ранньої зупинки роботи (зادля уникнення можливого перенавчання), метод Adam для оптимізації.

З результатами роботи початкових нейромереж можна ознайомитися нижче (рис. 3.52):

| Method                      | Quadratic Kappa |
|-----------------------------|-----------------|
| Regressor (Bins from PCIAT) | 0.369           |
| Classifier (Predict SII)    | 0.338           |

Рисунок 3.52 – Результати базових архітектур нейромереж

Як можна побачити з результатів, базові архітектури нейромереж видали досить посередні результати, порівняно з минулими етапами.

Задля можливого покращення значення коефіцієнту згоди Каппа, було вирішено провести роботи зі створення більш потужних архітектур початкових нейромереж. Таким чином, було реалізовано нейромережу, з додаванням декількох прихованих шарів, а також блоками зі skip-зв'язків.

Всі блоки нейромережі включають в себе два повнозв'язні шари з нормалізацією по батчах, функцією активації LeakyReLU (negative slope = 0.1) та методом Dropout для зменшення ймовірності перенавчання та

регуляризації. Архітектура на виході містить в собі три вищеописаних блоки з кількістю нейронів 1024, 512, 256 по шарам.

Архітектура просунутої нейронної мережі наведена нижче (рис. 3.53 – рис. 3.54):

| Структура ResidualBlock |                                              |
|-------------------------|----------------------------------------------|
| Компонент               | Конфігурація                                 |
| Linear 1                | Linear (in_dim $\rightarrow$ out_dim)        |
| BatchNorm 1             | BatchNorm1d (out_dim)                        |
| Activation 1            | LeakyReLU (0.1)                              |
| Dropout                 | Dropout (0.4)                                |
| Linear 2                | Linear (out_dim $\rightarrow$ out_dim)       |
| BatchNorm 2             | BatchNorm1d(out_dim)                         |
| Skip-зв'язок            | Identity, якщо in_dim == out_dim             |
|                         | Інакше Linear (in_dim $\rightarrow$ out_dim) |
| Вихід                   | LeakyReLU (x_main + x_skip)                  |

Рисунок 3.53 – Структура додаткового ResidualBlock

| Просунута нейромережа |                                                                                                                          |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------|
| Рівень                | Конфігурація                                                                                                             |
| Вхід                  | input_dim = N                                                                                                            |
| Вхідний блок          | Linear (N $\rightarrow$ 1024) $\rightarrow$ BatchNorm1d (1024) $\rightarrow$ LeakyReLU (0.1) $\rightarrow$ Dropout (0.5) |
| ResidualBlock 1       | ResidualBlock (1024 $\rightarrow$ 1024, dropout=0.5)                                                                     |
| ResidualBlock 2       | ResidualBlock (1024 $\rightarrow$ 512, dropout=0.5)                                                                      |
| ResidualBlock 3       | ResidualBlock (512 $\rightarrow$ 256, dropout=0.5)                                                                       |
| Вихід                 | Linear (256 $\rightarrow$ 1) / Linear (256 $\rightarrow$ 4)                                                              |

Рисунок 3.54 – Архітектура просунутої нейронної мережі

Класифікаційна нейронна мережа додатково використовує функцію втрат Focal Loss (gamma = 2.0) з label smoothing (0.05) для зниження впливу неякісних даних та фокусуванню на важко прогнозованих екземплярах, а також вагові коефіцієнти класів категоріальної змінної, що урівноважує їх дисбаланс.

Регресійна нейронна мережа як функцію втрат використовує MSELoss, оптимізатор AdamW, ранню зупинку навчання та обрізання градієнтів для

отримання більшої стабільності. Також було використано bagging-усереднення на основі трьох фіксованих чисел (посівів).

Як і в випадку з моделями класифікації та регресії, було вирішено додати окремим пунктом soft-ансамбль, що поєднує в собі обидві нейромережі з зваженим середнім їх прогнозів на виході. З результатами даного кроку роботи можна ознайомитися нижче (рис. 3.55):

| Method                                 | Quadratic Kappa |
|----------------------------------------|-----------------|
| Regressor (Bins from PCIAT)            | 0.378           |
| Classifier (Predict SII)               | 0.367           |
| Soft Ensemble (Regressor + Classifier) | 0.392           |

Рисунок 3.55 – Результати покращення базових архітектур нейромереж та їх поєднання в ансамбль

За результатами, як просунуті окремі нейронні мережі, так і фінальна архітектура з використанням ансамблю з двох мереж, на виході не продемонструвала значних результатів коефіцієнта згоди Каппа між отриманими результатами та реальними значеннями.

Внаслідок проведеної роботи з нейронними мережами, доречність їх використання в межах поставленої задачі прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників визнана недоцільною на цьому етапі результати. Однак не можна стверджувати, що використання нейромереж не має перспектив в досліджуваному напрямі, і наведені реалізації, можуть стати підґрунтям для подальших оптимізацій та робіт.

### 3.3.5. Фінальна модель

Підсумовуючи результати за трьома проведеними етапами експериментів та, було сформовано фінальну модель, яка дає найбільш точні та стабільні результати. Виконаний аналіз дав остаточне бачення, що

найкращою виявилась оптимізована модель CatBoost з підібраними гіперпараметрами:

1. Твідді-подібна функція втрат (loss\_function) (Tweedie:variance\_power=1.5), яка показала ефективну роботу на зашумлених біоповедінкових даних з асиметрією розподілу та неоднорідною структурою характеристик;
2. Кількість дерев (iterations) – 295, на яких модель не перенавчається і досягає найбільшої точності роботи;
3. Швидкість навчання (learning\_rate) – 0.04642, на якій досягалась стабільність та плавність роботи;
4. Глибина (depth) – 4, на якій було отримано можливість регуляції процесу узагальнення та запобігання перенавчанню;
5. Стохастичність (subsample) – 0.63193, за використанням якої модель набула більшої стійкості до міжознакової кореляції;
6. Регуляція випадковості розбиттів (random\_strength) – 7.1025, яка допомогла підвищити адаптивність моделі до неоднорідності даних;
7. L2-регуляризація (l2\_leaf\_reg) – 0.004, на якій стабільність дерев залишалась стало високою.

Фінальна архітектура регресійної моделі CatBoost (рис. 3.56):

```
cat_model = CatBoostRegressor(
    loss_function="Tweedie:variance_power=1.5",
    iterations=295,
    learning_rate=0.04642,
    depth=4,
    random_seed=42,
    subsample=0.63193,
    random_strength=7.1025,
    l2_leaf_reg=0.004,
    verbose=0
)
```

Рисунок 3.56 – Архітектура фінальної моделі

На етапі фінальної перевірки гіперпараметрів та тестування моделі на відповідній підвибірці даних було отримано наступні результати (рис. 3.57):

| Metric          | Value On Test |
|-----------------|---------------|
| Quadratic Kappa | 0.426         |

Рисунок 3.57 – Фінальна оцінка на етапі тестування

Обрана модель показала найвищу узгодженість спрогнозованих значень з реальними, на основі коефіцієнта згоди Каппа, що аргументує доцільність саме такої стратегії подальшої роботи для поставленої задачі прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників.

### Висновки до розділу 3

Цей розділ містить в собі наймасштабнішу частину дослідження – вибір необхідного технологічного стеку, датасету, роботи з ним та методами машинного навчання, визначення фінальної моделі. Основними етапами є:

1. Вибір технологічного стеку, в результаті якого було прийнято рішення про використання мови програмування – Python; бібліотек – NumPy, Pandas, Tabulate, Scikit-learn, PyTorch, Skorch, XGBoost, LightGBM, CatBoost, Matplotlib, Seaborn, Tkinter, Joblib; середовища розробки – PyCharm.

2. Вибір та робота з датасетом, що включила в себе огляд набору даних, роботу з різними видами даних, обробку пропущених значень, маніпуляції зі списком ознак та їх важливістю;

3. Вибір та створення моделі машинного навчання – опис загальної схеми роботи, аналіз класифікаційних, регресійних та нейромережових моделей, підбір гіперпараметрів для них, створення ансамблів різних типів, та, у результаті, отримання фінальної моделі CatBoost, що показала найбільші значення коефіцієнту згоди Каппа між спрогнозованими значеннями та реальними даними.

## **РОЗДІЛ 4**

### **ПРОГРАМНА РЕАЛІЗАЦІЯ ТА МЕТОДИКА РОБОТИ СИСТЕМИ ПРОГНОЗУВАННЯ РІВНЯ ПРОБЛЕМНОГО ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ**

В цьому розділі МД розглядатиметься практичне створення програмного забезпечення з візуальним інтерфейсом користувача. В основу додатку буде включено обрану та навчену фінальну модель машинного навчання. Також описано функціональні вимоги до застосунку, відображено кроки його реалізації та показано методику роботи.

#### **4.1. Функціональні вимоги до програмного застосунку**

Функціональні вимоги – це вимоги до можливостей програмного застосунку, набір дій, які він зможе виконувати, включаючи методику його роботи та взаємодії, очікувані результати тощо.

Таким чином, нижче описані функціональні вимоги до програмного застосунку з прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників:

1. Повідомлення користувача про політику конфіденційності під час роботи додатку;
2. Можливість імпорту вхідних даних в форматі CSV, які являють собою оцінки за результатами психологічних анкет і опитувальників, фізіологічні параметри та інші дані;
3. Можливість ручного введення вхідних даних в окремому вікні застосунку;
4. Перевірка коректності вхідних даних (перевірка наявності даних у файлі та коректність їх заповнення, перевірка конкретних значень – числовий формат, додатність тощо), та, за потреби, повідомлення користувача про помилку в додатковому вікні;

5. Заповнення пропущених значень за допомогою завчасно навченого та збереженого імпутера;
5. Масштабування і нормалізація змінних;
6. Виконання прогнозу рівня проблемного використання Інтернету за вхідними даними, використовуючи завчасно навчену та збережену модель CatBoost із підібраними гіперпараметрами;
7. Візуалізація отриманих результатів в окремому вікні;
8. Інтерфейс програмного застосунку – зручний, легкий для розуміння й використання, як для фахівців, так і для вчителів та батьків.

## **4.2. Програмна реалізація програмного застосунку**

ПЗ «Problematic Internet Use Predictor» побудовано як десктопну програму, написану на мові програмування Python, з графічним інтерфейсом користувача з використанням бібліотеки Tkinter та із застосуванням фінальної моделі машинного навчання для прогнозування результируючих значень та виведенням цих результатів користувачу.

Додаток пропонує два варіанти внесення даних (через файл або вручну), перевірку коректності введеної інформації, безпечну роботу з пропущеними ознаками (які не є ключовими) та проведення прогнозування рівня проблемного використання Інтернету дітьми за допомогою попередньо навченої та збереженої моделі.

### **4.2.1. Архітектура програмного застосунку**

Застосунок являє собою одновіконний графічний інтерфейс користувача з декількома модальними вікнами. Роботу програми можна розділити на наступні логічні блоки:

1. Робота з артефактами машинного навчання – у момент запуску застосунку з папки models зчитується файл catboost\_regressor.pkl, в якому міститься словник з даними model, imputer, predictors. ІМодель – регресор CatBoost з підібраними та оптимізованими гіперпараметрами.

2. Диспетчер ознак – список під назвою `all_features`, що містить набір ознак, які очікуються, а також їх порядок; список під назвою `feature_labels`, що містить зрозумілий для користувача опис ознак, а також повідомлення для валідації.

3. Модуль валідації вхідних даних – система правил `validation_rules` і `required_fields`, а саме: перевірка типів даних (тільки числові значення), перевірка можливих, завчасно визначених для дітей діапазонів значень для кожної з ознак, перевірка від’ємних та логічно неможливих значень, визначення обов’язкових ознак (найбільш важливих), без яких процес прогнозування не може бути ініційовано.

4. Обробка пропущених значень – за наявності відповідного імпутера даних із артефакту зі збереженням порядку ознак, інакше – заповнення середніми значеннями.

5. Процес прогнозування та класифікація результату – дані подаються на вхід до зчитаної моделі машинного навчання, яка на виході видає значення оцінки PCІАТ, який далі розподіляється за чотирма класами за визначеними порогоми, залежно від рівня проблемності використання мережі Інтернет – None, Mild, Moderate, Severe. Фінальний результат супроводжується коротким описом та рекомендацією, що виводиться в додатковому спливаючому вікні.

6. Рівень графічного інтерфейсу – дисклеймер конфіденційності, передує всім іншим діям в застосунку; головне вікно застосунку; форма для ручного введення даних. Інтерфейс виконано в спільному для всього додатку стилі – шрифт Segoe UI, використовуються закруглені рамки, кнопки, що вирізняються контрастністю, підсвітка `hover state`. Застосовуються модальні вікна для додаткових повідомлень.

#### **4.2.2. Потік даних під час роботи програмного застосунку**

1. Введення даних – ручне, або зчитування з CSV-файлу.

2. Перевірка даних – приймаються тільки файли з одним рядком даних; порівняння заголовків файлу зі списком ознак, що очікує модель на вхід;

нечислові рядки конвертуються (або видається помилка за неможливості); ознаки, які не вдалося розпізнати, опускаються. При будь-яких некоректно введених даних користувача повідомляється про виникнення помилки.

3. Валідація даних – перевірка діапазонів значень, чи є всі обов’язкові ознаки.

4. Відновлення або заповнення пропущених даних.

5. Прогнозування результатів, переведення значення у класову мітку.

6. Виведення повідомлення про отримані результати та рекомендації.

#### **4.2.3. Тестування роботи програмного застосунку**

Тестування роботи програмного застосунку включало в себе:

1. коректна робота програми;

2. некоректна робота програми, зокрема: порожній файл, кількість рядків більше одного, відсутність обов’язкових для процесу прогнозування ознак, введення тексту в числові комірки, кома як роздільник в значеннях із десятковою точністю, значення поза встановлених діапазонів;

3. стабільність заповнення пропущених значень та фіксація порядку ознак;

4. консистентність отримуваних прогнозів за незмінності вхідних параметрів.

#### **4.3. Методика роботи програмного застосунку**

Запуск програмного застосунку супроводжується спливаючим вікном, що повідомляє користувача про те, що всі дані, які він вносить при користуванні додатком ніде не зберігаються, використовуються лише локально, під час поточної сесії, після чого вони видаляються автоматично. Таким чином, користувач може або погодитись, або вийти із застосунку (рис. 4.1):

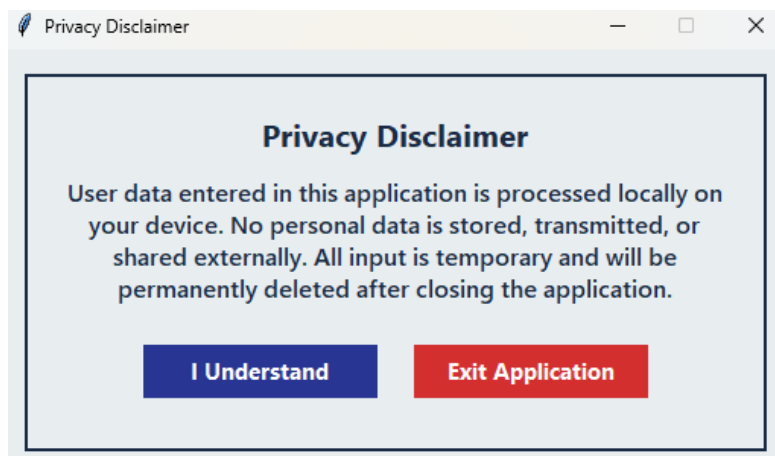


Рисунок 4.1 – Повідомлення про політику конфіденційності

Якщо користувач погоджується, спливаюче вікно закривається, після чого відкривається основне вікно додатку, з двома опціями для роботи – вибір CSV файлу, вже наповненого всіма необхідними даними, або ж введення даних вручну (рис. 4.2):

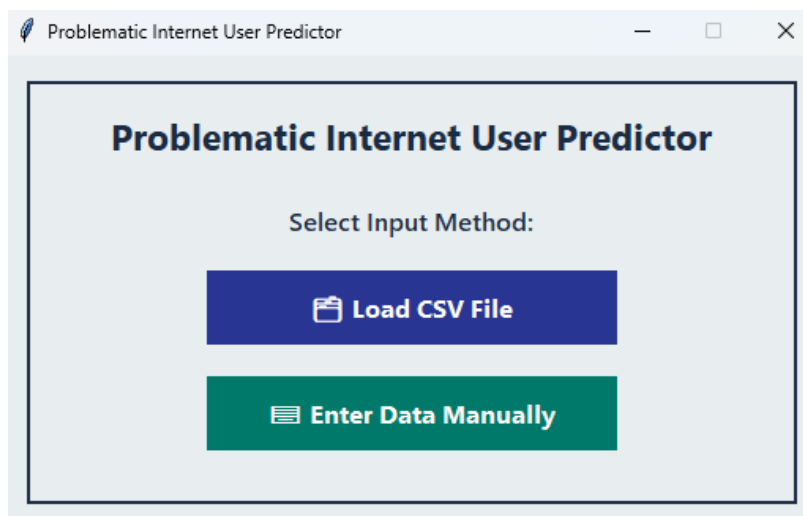


Рисунок 4.2 – Головне вікно програмного застосунку

При виборі опції зчитування даних з CSV-файлу відкривається вікно файлового провідника, в якому користувач має вибрати свій файл з даними, на основі яких буде виконуватися прогнозування (рис. 4.3):

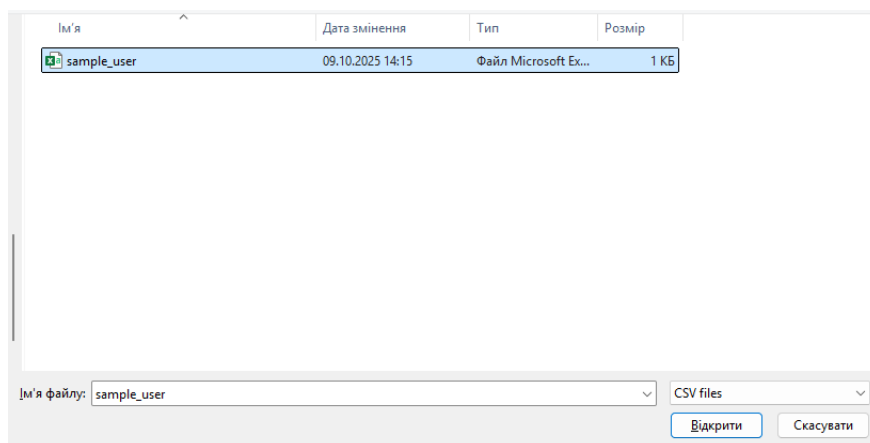


Рисунок 4.3 – Відкриття файлового провідника для вибору CSV файлу

Після вибору файлу відбувається зчитування та перевірка даних:

1. Дані мають бути числовими, і не містити від'ємних значення;
2. Дані мають бути у логічних межах та з урахуванням, що прогнозування йде для дитини (таким чином були проставлені числові обмеження для вхідних параметрів);
3. Дані обов'язково мають містити попередньо визначений набір параметрів (найбільш корельовані та логічно обґрунтовані ознаки).

Якщо файл було створено та заповнено вірно, він не містив некоректно введених або відсутніх обов'язкових даних, запускається процес прогнозування, та перед користувачем виводиться передбачений рівень проблемного використання Інтернету його дитиною та рекомендація щодо подальших дій (рис. 4.4):

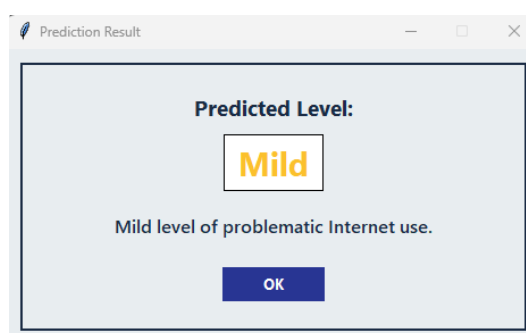


Рисунок 4.4 – Вікно результатів прогнозування

При виборі некоректно складеного файлу, відображається спливаюче вікно з текстом помилки (рис. 4.5):

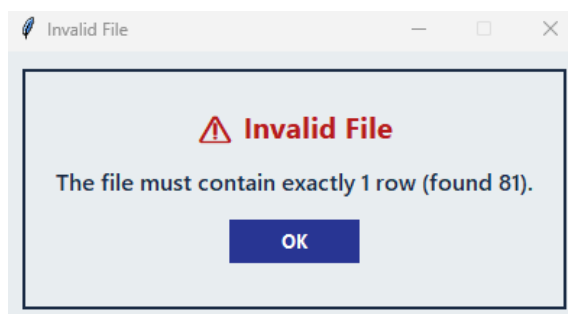


Рисунок 4.5 – Помилка зчитування файлу

За відсутності обов’язкових полів відображається відповідне вікно з помилкою (рис. 4.6):

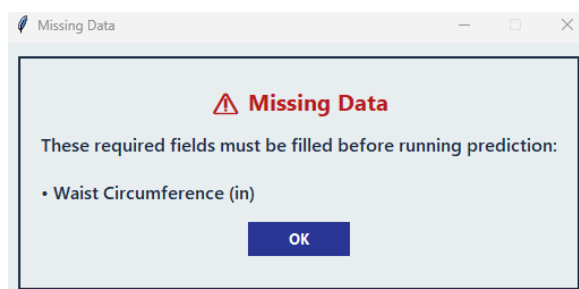


Рисунок 4.6 – Відсутність обов’язкових ознак вхідних даних

При некоректно заповнених даних, відображається відповідні вікна з помилками (рис. 4.7 – рис. 4.8):

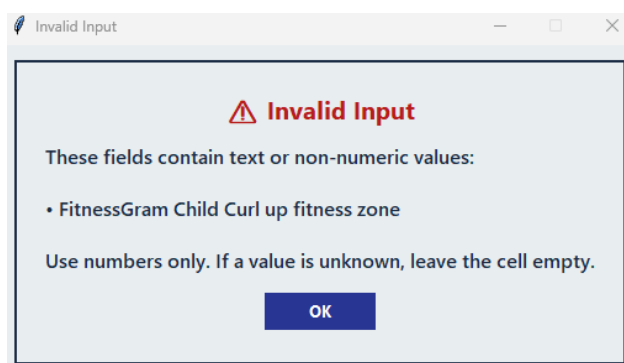


Рисунок 4.7 – Нечислові значення вхідних даних

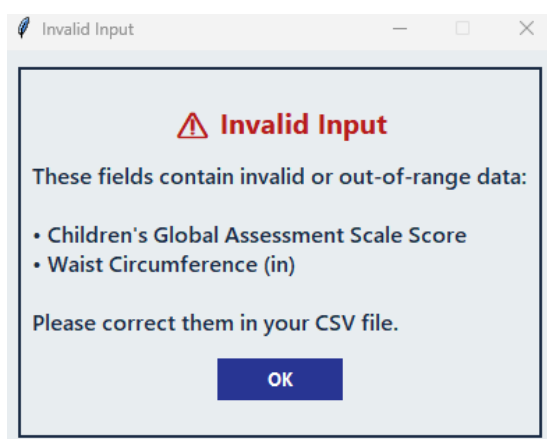


Рисунок 4.8 – Некоректні значення вхідних даних

При виборі опції ручного заповнення даних перед користувачем відкривається вікно програми з відповідними полями для внесення інформації (обов'язкові поля позначаються знаком зірочки «\*») (рис. 4.9):

Рисунок 4.9 – Вікно вручного введення даних

Після введення даних користувачем, він має натиснути кнопку для запуску процесу передбачення. Відбувається перевірка вхідних даних, імпуція пропущених значень (якщо ознака не позначена обов'язковою) та проведення прогнозування. При коректному введенні інформації перед користувачем виводиться передбачений рівень проблемного використання Інтернету дитиною та рекомендація щодо подальших дій, а при невірному – відповідні повідомлення.

#### **Висновки до розділу 4**

Цей розділ дослідження показав практичну реалізацію системи прогнозування рівня проблемного використання Інтернету дітьми.

ПЗ було узгоджено зі всіма функціональними вимогами, які передбачають наявність графічного інтерфейсу, можливість введення, перевірки та обробки даних, а також передавання їх в модель, що на виході формує передбачення. Модульність архітектури застосунку гарантувала надійність та зрозумілість роботи програми, а проведене тестування підтвердило коректність і якість його роботи.

## **РОЗДІЛ 5**

### **УЗАГАЛЬНЕННЯ РЕЗУЛЬТАТІВ РОБОТИ ЗІ СТВОРЕННЯ СИСТЕМИ ПРОГНОЗУВАННЯ РІВНЯ ПРОБЛЕМНОГО ВИКОРИСТАННЯ ІНТЕРНЕТУ ДІТЬМИ**

Цей розділ містить оглядовий опис всієї проведеної роботи, в якому підведено підсумки за всіма етапами дослідження по створенню спеціалізованої системи прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників.

#### **5.1. Аналіз отриманих результатів**

В результаті проведеного дослідження, було здобуто комплексні результати, що вмістили в собі огляд та обґрунтування проблеми з точки зору теорії, створення узагальненої структури даних, побудову архітектури роботи та вибір найкращого рішення машинного навчання, а також створення програмного застосунку, що функціонує в якості повноцінного додатку для прогнозування рівня проблемного використання Інтернету дітьми.

У межах теоретичного етапу було проведено аналіз наявних на цей момент методів та підходів до виявлення Інтернет-залежності, доведено важливість використання саме комбінації фізіологічних, психоемоційних, поведінкових та інших даних. Таким чином була сформована основна концептуальна модель проведення дослідження.

Вибір технологічного стека проводився спираючись на вимоги до якості, точності, ресурсів системи, та виходячи з поставленої задачі. В результаті, було обґрунтовано вибір мови програмування Python, його бібліотек, середовища розробки PyCharm.

За набір даних було взято датасет «Problematic Internet Use» від Child Mind Institute, що знаходиться в відкритому доступі, та який поєднує велику кількість різноманітних даних щодо дітей та підлітків (фізіологічні та

біоімпедансні показники, результати роботи актиграфа, результати психологічних опитувальників тощо). Дані було очищено, приведено до спільного формату, оброблено пропущені значення. На основі набору було сформовано вибірку даних, що поєднала всі ознаки, додано нові та видалено неякісні або непотрібні. Результуюча вибірка надавалася напряду до моделі машинного навчання.

Етап аналітики вказав на наявність кореляцій між фізичними даними та ключовими характеристиками, які прогнозуються, демонструючи рівень проблемного використання Інтернету, що підтвердило можливість використання такого характеру даних в якості індикаторів залежності від мережі, тобто актуальності дослідження.

В результаті проведення моделювання, було протестовано роботу декількох алгоритмів машинного навчання різних типів, нейронних мереж, а також їх оптимізованих версій. В результаті, найкращий результат показав регресор градієнтного бустингу CatBoost, а також було підтверджено його узагальнювальну властивість та відсутність перенавчання за допомогою кросвалідації.

Останнім кроком практичної реалізації системи стала розробка програмного застосунку за завчасно визначеними функціональними вимогами, за допомогою якого створюється можливість взаємодії користувача з фінальною прогнозувальною моделлю. Було реалізовано валідацію даних, що надходять, заповнення пропущених значень ознак (які не є обов'язковими), класифікацію рівня проблемного використання Інтернету дитиною та візуалізацію отримуваних результатів через інтерфейс користувача (побудованого з урахуванням правил доступності, зручності та безпеки).

Підсумовуючи, отриманий застосунок являє собою цілісну та ефективну в роботі систему – скомбінувавши аналітичний підхід та теоретичне обґрунтування алгоритмічної основи та практичної реалізації було створено потужний інструмент з прогнозування рівня проблемного використання Інтернету дітьми на основі комплексу різноманітних, зокрема фізичних даних,

який в подальшому має перспективи широкого застосування в сферах освіти, психології, медицини тощо. За допомогою такого рішення, відсоток молоді, залежної від мережі Інтернет може значно понизитися, тим самим покращуючи якість їх життя та здоров'я.

## **5.2. Унікальність проведеного дослідження**

### **5.2.1. Схема роботи з даними**

Першочергово, дане дослідження, вирізняється своєю багаторівневою, унікальною роботою над формуванням вхідних вибірок даних на вхід до моделей машинного навчання. За відправну точку було взято великий датасет від Child Mind Institute під назвою «Problematic Internet Use». Він вмістив у собі фізіологічні та біоімпедансні показники, дані фітнес-тестів та різних психоемоційних опитувальників, записи роботи актиграфу тощо.

Схема роботи з даними охоплює наступні етапи:

1. Базовим етапом є завантаження даних зі всіх наявних файлів, їх об'єднання за допомогою унікального ідентифікатора кожного досліджуваного, приведення до спільної структури.

2. Робота з даними актиграфії – комплексний препроцесинг: зведення до єдиного формату часових міток та значень сигналів; фільтрація невалідних сегментів; початкове очищення зашумлених даних; визначення додаткових статистик; додавання агрегованих характеристик щодо активності.

3. Інша частина даних – табличні характеристики, опрацьовувались наступним чином: зведення до єдиного формату; типізація даних, приведення; початкова робота з нецілісними ознаками; кодування класифікаційних змінних.

4. Відновлення та заповнення пропущених значень з урахуванням структурних особливостей та топології даних. Цільова ознака оцінки RCIAT-тестування заповнювалась за допомогою методу MICE, в основі якого покладено алгоритм випадкового лісу з подальшим перерахуванням класу рівня проблемності використання мережі. Усі інші залежні ознаки,

заповнювались методом множинної імпутації, що відновлює пропущені дані через послідовну побудову регресійних рівнянь між змінними.

5. Створення додаткових змінних – у вже об'єднаній структурі, що містила всі дані додавалися нові ознаки, за допомогою яких розширювалася інформативність вибірки даних.

6. Видалення надлишкових ознак було проведено виходячи з результатів кореляційного аналізу, а також логічних зв'язків між характеристиками. Таким чином, повторювані, колінеарні, неповноцінні, змінні, які не несли важливої інформації було вилучено, тим самим досягаючи стабільності подальшої роботи.

7. Визначення набору найбільш важливих ознак для подальшого акцентування уваги саме на них.

8. Опціональна нормалізація значень для чутливих до цього моделей.

9. Формування фінальних вибірок даних, готових для подання на вхід до моделей машинного навчання.

### 5.2.2. Фінальна конфігурація моделі

За фінальну модель було обрано регресор градієнтного бустингу CatBoost, різні конфігурації якого розглядалися під час тестування роботи алгоритмів машинного навчання, унаслідок чого було знайдено оптимальні значення гіперпараметрів.

Таким чином, фінальна оптимізована модель (рис. 5.1):

```
cat_boost_model = CatBoostRegressor(  
    loss_function="Tweedie:variance_power=1.5",  
    iterations=295,  
    learning_rate=0.04642,  
    depth=4,  
    random_seed=42,  
    subsample=0.63193,  
    random_strength=7.1025,  
    l2_leaf_reg=0.004,  
    verbose=0  
)
```

Рисунок 5.1 – Фінальна конфігурація моделі

### **5.2.3. Порівняння результатів роботи моделі з наявними дослідженнями**

Унаслідок проведеної роботи, було отримано фінальну конфігурацію моделі з підібраними гіперпараметрами, та протестовано її на тестовій вибірці даних, оцінюючи на основі значення метрики коефіцієнта згоди Каппа, що є найбільш об'єктивною та інформативною в рамках дослідження.

Таким чином, було отримане значення помірною рівня узгодженості прогнозованих значень з реальними даними (0.426). Порівнюючи зі всіма іншими, наявними роботами, що проводились з регресором градієнтного бустингу CatBoost, отриманий результат перевищує найкращі значення (0.4086) на 4.25 відсотка, що є досить гарним результатом.

Підсумовуючи, запропонована конфігурація моделі регресора CatBoost перевищила значення показників минулих конфігурацій, забезпечуючи більшу якість та точність отримуваних результатів з прогнозування рівня проблемного використання Інтернету дітьми.

## **5.3. Практичне значення результатів**

Створена система прогнозування рівня проблемного використання Інтернету дітьми, що використовує фізичні, та інші дані, має перспективу використання в багатьох сферах та відкриває широкі можливості.

Основними напрямками подальшого застосування системи є:

1. Медицина та психологія – розроблена система може знайти своє використання як у клінічних, так і закладах консультаційного характеру, в якості спеціалізованого інструмента для об'єктивної оцінки рівня проблемності використання мережі та ризику формування Інтернет-залежності для сімейних лікарів та неврологів, психологів та психіатрів. Перехід від більш суб'єктивно працюючих психометричних тестів до науково обґрунтованого алгоритму створює можливість надання більш

індивідуального підходу до профілактики та раннього виявлення психологічних та емоційних порушень у молоді.

2. Освіта та соціальне середовище – розроблена система в подальшому може бути інтегрована в служби шкільної інфраструктури та освітні платформи та сервіси, пропонуючи батькам та педагогам оперативні аналітичні дані відносно стану цифрового благополуччя дитини чи підлітка. Таким чином можна буде досягти значного покращення інформаційної безпеки та методів моніторингу без порушення конфіденційності.

3. Система як готовий продукт – ПЗ наразі вже має свій власний інтерфейс користувача, проводить повноцінний цикл роботи з даними, а також забезпечує виведення рівня проблемного використання Інтернету та короткі рекомендації. Одним з варіантів подальшого розвитку є доопрацювання вже наявної системи та використання її як готового продукту.

## **Висновки до розділу 5**

Цей розділ узагальнив всі отримані результати протягом дослідження та підтвердив ефективність застосування фізичних даних в якості індикаторів проблемного використання Інтернету дітьми та визначення його рівня. Поєднання таких фізіологічних параметрів з іншими даними, такими як демографічні, актиграфічні, психоемоційні, когнітивні, поведінкові ознаки, дало змогу створити високоякісну та точну, комбіновану систему зі здатністю до раннього виявлення ризику виникнення цифрової залежності.

Під час дослідження, було проведено повноцінний цикл обробки даних – очищення, інженерія ознак, заповнення пропусків, видалення невалідних та визначення найважливіших змінних тощо. Після цього проводилась робота з методами та алгоритмами машинного навчання, внаслідок чого для поставленої задачі прогнозування в якості фінальної моделі було обрано регресійний метод градієнтного бустингу CatBoost, що на виході

демонстрував найкращу стабільність результатів та найбільші значення метрики оцінки якості – коефіцієнта згоди Каппа.

Отримана модель машинного навчання була використана в програмному застосунку, що має унікальний та зручний графічний інтерфейс, з можливістю зчитування даних дитини через відкриття файлу або ж ручного введення у вікні програми, після чого відбувається процес прогнозування, виведення результатів та рекомендацій. Така система має перспективи практичного використання в медичних, психологічних та освітніх організаціях та навіть застосовуватися як готовий продукт.

Таким чином, результати проведеного дослідження підтверджують його актуальність, соціальну важливість, а також успішність підходу, що розглядався – поєднання методів машинного навчання у тандемі з різноманітними видами даних, та, зокрема, фізичних. Все це вказує на можливість широкого використання створеної системи в найближчому майбутньому для раннього виявлення та профілактики проблемного використання Інтернету в молоді.

## **РОЗДІЛ 6**

### **СТАРТАП-ПРОЄКТ ЗА ТЕМОЮ МАГІСТЕРСЬКОГО ДОСЛІДЖЕННЯ**

Цей розділ присвячений розробці стартап-проєкту за темою МД – створення системи прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників. Було наведено концепцію створюваного продукту, його організацію, досліджено ринок, описано бізнес-модель, види діяльності, ключові ресурси та партнерів, фінансовий план та приблизні витрати, канали збуту та отримання доходу, проведено аналіз цільової аудиторії та взаємодії з нею.

#### **6.1. Короткий опис проєкту**

Стартап-проєкт по створенню програмного застосунку прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників.

Проєкт спрямовано на реалізацію відповідної системи, яка, за допомогою фізичних (в основі) та статистичних показників, матиме можливість якісного та швидкого прогнозування рівня проблемного використання Інтернету дітьми. Основною ідеєю є використання методів машинного навчання та нейронних мереж для аналізу наявних даних (показників тіла, рівня фізичної активності, даних щодо сну, часу використання мережі, результатів опитувальників тощо) для використання в якості підходу до раннього виявлення прихильності дитини до неконтрольованого та надмірного застосування Інтернету, з подальшим втручанням, з метою покращення життєдіяльності та майбутнього життя молоді.

Головними особливостями стартап-проєкту можна назвати такі:

1. Комбінація фізичних даних з класичними анкетами й опитувальниками та використанням сучасних технологій;
2. Використання індивідуальних даних кожного користувача;
3. Точність та якість отримуваних прогнозованих значень.

Цей стартап-проєкт охоплює реалізацію зручної, як для лікарів та психологів, так і для батьків та вчителів системи, яка міститиме функціонал з виявлення Інтернет-залежності за різними рівнями її вираженості, та надаватиме можливість вчасного втручання та задіяння необхідних профілактичних методів.

ПЗ вартий уваги користувачів, які хочуть покращити психічний та емоційний стани своїх дітей, учнів або пацієнтів, мінімізувати ризики виникнення негативних фізичних або соціальних наслідків використання мережі Інтернет, формуючи повноцінне життя молоді в добу цифрових технологій.

## **6.2. Бізнес-модель**

### **6.2.1. Цінність та вимоги до продукту**

Ключовою цінністю стартап-проєкту є його інноваційність в Україні (потенційно у всьому світі), а саме системи, що комбінує різноманітні види даних (фізичні параметри, статистичні ознаки, результати психологічних анкет та опитувальників) та новітні методи і підходи, в задачі прогнозування рівня проблемного використання Інтернету дітьми. Порівняльний аналіз усіх можливих конкурентних рішень в обраній області можна побачити нижче (рис. 6.1).

| Назва конкурента           | Короткий опис діяльності                                                | Короткий опис продукту                                                                 | Схожість зі стартап-проектом                                        |
|----------------------------|-------------------------------------------------------------------------|----------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| Bark Technologies          | Платформа для моніторингу онлайн-активності дітей                       | Аналіз повідомлень, пошуку, YouTube тощо; попередження про ризики (булінг, залежність) | Висока: виявляє ризики, але орієнтована на контент, а не фізіологію |
| Qustodio                   | Контроль цифрової поведінки дитини                                      | Обмеження часу на пристрої, контроль активності, аналітика                             | Середня: фокус на часі в онлайні, а не на внутрішніх причинах       |
| Replika.ai                 | Емоційний AI-компаньйон                                                 | Платформа для симуляції розмов, використовується підлітками, має ризики залежності     | Низька: потенційний конкурент лише як фактор ризику                 |
| BetterHelp (parent access) | Онлайн-платформа психологічної підтримки                                | Доступ до психотерапевтів для батьків і підлітків онлайн                               | Середня: надає підтримку після виявлення проблеми                   |
| Шкільний психолог (офлайн) | Державна/муніципальна послуга з виявлення поведінкових ризиків в школах | Збір скарг, бесіди з дітьми, фрагментарна діагностика                                  | Середня: залежить від людського чинника, неавтоматизована           |

Рисунок 6.1 – Огляд конкурентних рішень

Порівняння з вже наявними рішеннями, які в основному фокусуються на моніторингу поведінки дитини (статистичні показники використання Інтернету, контент, який поглинає користувач), показало – створювана система вирізняється такими унікальними характеристиками:

1. Відсутність необхідності прямого втручання до конфіденційної цифрової інформації дитини;
2. Використання фізичних показників як можливих маркерів залежності;
3. Точність та якість, що досягається за допомогою алгоритмів машинного навчання та нейронних мереж;
4. Можливість інтеграції до навчальних, медичних платформ тощо.

Будь-який ПП має підпорядковуватися певному набору функціональних та нефункціональних вимог.

Функціональні вимоги:

1. Збір, аналіз, обробка та подання на вхід даних;
2. Проведення прогнозування рівня проблемного використання Інтернету;
3. За результатами надання користувачу персонального ризик-профілю;

4. Виведення рекомендацій щодо подальших дій.

Нефункціональні вимоги:

1. Конфіденційність та захищеність використовуваної інформації;
2. Інтегрованість до різноманітних платформ та пристроїв (в майбутньому);
3. Масштабованість реалізованої системи, можливість її локалізації, підлаштування до різних груп людей.

### **6.2.2. Сегмент споживачів**

Основними споживачами продукту, попередньо було визначено батьків дітей, освітні та медичні установи (та їхні працівники), які є зацікавленими в ранньому виявленні та профілактиці розладів, пов'язаних з проблемним використанням мережі Інтернет серед молодого покоління.

Основними причинами актуальності серед споживачів можна назвати:

1. Часта тривожність серед батьків, що виникає внаслідок надмірного використання інтернету дітьми;
2. Поточна відсутність конкуренції серед аналогічно працюючих засобів, які б не впливали на приватність дитини;
3. Нерозвиненість використання різних видів даних (таких як фізичні), під час встановлення діагнозів, які, в свою чергу, можуть бути надійними індикаторами цифрової залежності.

Поєднувальна мотивація споживачів:

1. Прагнення захисту дітей від можливих проблем та загроз;
2. Зростання необхідності сучасного та наукового підходу до обґрунтування отримуваних результатів;
3. Усвідомлення актуальності тематики та можливих ризиків, пов'язаних з надмірною та неконтрольованою активністю в мережі Інтернет.

Модель поведінки споживача:

1. Застосування носимих пристроїв для моніторингу здоров'я;
2. Звернення до спеціалістів за наявності ознак Інтернет-залежності;
3. Встановлення цифрового контролю.

Тип сегменту споживачів:

1. B2C – безпосередньо батьки дитини, що матимуть пряму можливість встановлення та використання програмного застосунку.
2. B2B або B2G – навчальні заклади, медичні установи, приватні клініки, в якості користувачів, або проміжних партнерів для масового поширення та впровадження продукту.

### **6.2.3. Канали збуту**

Ключовими каналами збуту ПП будуть являтися наступні:

1. Прямі продажі за використанням офіційного вебсайту чи застосунку, таким чином для кожного типу сегментів споживачів:
  - батьки – довгострокова підписка або разове придбання доступу до системи;
  - установи – корпоративні підписки та ліцензії з доступом до сервісу, або індивідуальна інтеграція до вже наявних систем.
2. Непрямі продажі за використанням проміжних платформ:
  - через інтеграцію в наявні системи дистанційної освіти, медицини тощо;
  - через встановлення програмного застосунку в базовий функціонал носимих пристроїв та гаджетів.
3. Інституційні продажі (B2B або B2G):
  - за допомогою завчасно підготованих презентацій в партнерських установах;
  - за допомогою комбінування з державними та локальними програмами з дослідження психоемоційного стану молоді.

Підтримка споживачів після придбання та канали комунікації:

- інструкції з використання системи, розділ Frequently Asked Questions (FAQ) на вебсайті або в застосунку, спеціалізовані блоки інформації для фахівців;
- форум, канали в соцмережах для комунікації, обміну досвідом та взаємодопомоги;
- онлайн-консультація фахівців різного виду (від технічної підтримки до психологів та терапевтів);
- налаштовувані розсилки інформації щодо оновлень, порад, аналітичних даних та статистики.

#### **6.2.4. Взаємодія зі споживачами**

Етап залучення споживачів:

1. Реклама через соцмережі, направлена на цільового споживача;
2. Розміщення матеріалів за тематикою «Психологія», «Цифрове здоров'я»;
3. Рання «пілотна» інтеграція ПП у визначених установах;
4. Участь у різноманітних заходах з освітніх та наукових технологій тощо;
5. Реалізація реферальної програми.

Етап утримання споживачів:

1. Індивідуальні результати для кожного набору даних дитини та персоналізовані рекомендації;
2. Спеціалізовані повідомлення з приводу проведення повторної роботи з прогнозування, корекції рекомендацій;
3. Регулярне покращення функціоналу, оновлення та додавання корисних матеріалів.

Етап розвитку відносин зі споживачами:

1. Надання індивідуальних консультацій з реальними фахівцями (додаткова плата або преміум-підписка);
2. Розширений функціонал: аналітика даних, візуалізація результатів;
3. Додаткові інтеграції з різноманітними носимими пристроями;
4. Мультиакаунтинг для установ з корпоративним доступом.

#### **6.2.5. Дохід (монетизація)**

Стартап-проект оснований на гібридній моделі заробітку, що містить в собі заробіток з прямих підписок та ліцензій, інтеграційних співпраць та додаткових користувацьких послуг. Таким чином досягається масштабованість та адаптивність проекту. Нижче наведено детальний опис:

1. Прямі підписки для поодиноких користувачів, засновані на freemium-моделі – базовий безкоштовний доступ (обмежений доступ, кількість використань), платний додатковий функціонал за місячною, довгостроковою або довічною підпискою.
2. Корпоративний доступ та ліцензії для організацій (школи, медичні установи, клініки тощо) – доступ до системи через спеціалізовані пакети з можливістю аналітики даних.
3. Інтеграції та партнерства – створення API для швидкого та зручного використання партнерськими платформами, виробниками пристроїв. Дохід надходить через відсоток від продажів, ліцензії або за допомогою так званих white-label партнерств.
4. Додаткові послуги: консультації з фахівцями (дитячими психологами, педіатрами та терапевтами, технічними спеціалістами) через платформу системи за додаткову оплату.

### **6.2.6. Ключові види діяльності**

Ключовими видами діяльності можна назвати повний життєвий цикл ППІ від створення та впровадження, до подальшої підтримки системи прогнозування рівня проблемного використання Інтернету дітьми.

Нижче описано всі етапи та процеси:

1. Розробка програмного застосунку – десктопний, мобільний та вебзастосунок, з інтерактивним та зручним інтерфейсом користувача, правильно вибудованої бекенд-архітектури, якісними алгоритмами машинного навчання та нейронними мережами з обробки даних та подальшого прогнозування.

1.1. Збір та обробка даних різних видів та форматів – їх очищення від некоректної та зашумленої інформації, заповнення пропущених значень, зведення до єдиного спільного формату.

1.2. Створення моделей машинного навчання та нейромереж – розробка, тренування, налаштування, валідація прогнозувальних алгоритмів рівня Інтернет-залежності, на основі отриманих даних.

1.3. UI- та UX-дизайн та аналітика майбутніх користувачів – створення та реалізація інтерфейсу, що підійде для різних сегментів споживачів, на основі проведеного аналізу потреб.

2. Маркетинг – цифрове просування різноманітні співпраці, виступи на конференціях, публікації в журналах та освітніх документах, пошук партнерів для подальшої роботи.

3. Підтримка споживачів – на основі зворотного зв'язку від користувачів система може зазнавати змін для покращення своєї роботи, надання консультаційних послуг з різних питань.

4. Захист інформації та підпорядкування етичним нормам – будь-яка діяльність у межах стартап-проекту відповідає всім правовим нормам та законам, впровадження політики безпеки, захисту доступу до даних і персональної інформації.

### 6.2.7. Ключові ресурси

Для реалізації стартап-проєкту необхідно мати ресурси трьох ключових типів: людські ресурси, інтелектуальні ресурси та фінансові ресурси.

#### 1. Людські ресурси:

- керівник проєкту (стратегія та управління);
- full-stack розробник (front-end та back-end);
- інженер у сфері машинного навчання (розробка прогнозувальних моделей);
- UX та UI дизайнер (інтерфейс та візуальна частина системи);
- фахівець із безпеки інформації (опціонально в майбутньому);
- консультанти у сфері психології та медицини (дитячий психолог, педіатр);
- технічний фахівець для підтримки (оновлення платформи та системи, спілкування з користувачами).

#### 2. Інтелектуальні ресурси:

- створені, навчені та налаштовані моделі машинного навчання та нейронні мережі;
- інтерфейсна логіка та дизайн;
- методики та підходи до аналізу результатів прогнозування й оцінки ризикових станів, отримані внаслідок проведення експериментів та тестувань.

#### 3. Фінансові ресурси:

- витрати на створення мінімального життєздатного продукту (MVP) (грант або інвестиції);
- витрати на маркетинг та просування продукту;
- витрати на технічне обладнання та програмне забезпечення;
- витрати на оплату праці спеціалістів;
- витрати на юридичні послуги.

### **6.2.8. Ключові партнери**

Реалізація та впровадження стартап-проєкту напряду залежить від ключових партнерів, що своєю участю можуть значно посприяти розширенню функціональних можливостей системи, надати доступ до цільових споживачів та позитивно впливати на рейтинг проєкту серед населення.

Ключовими партнерами проєкту можуть бути:

1. Освітні установи (дитсадки, школи тощо) – можливість тестування системи в реальних умовах, отримання відгуків безпосередньо від батьків та педагогів.
2. Медичні установи (лікарні, центри здоров'я, приватні клініки) – можливість отримання експертної думки, консультацій зі спеціалістами, формування порад та рекомендацій для користувачів.
3. Виробники розумних гаджетів та носимих пристроїв – можливість інтеграції з цими самими пристроями, підвищення популярності використання системи та полегшення збору інформації щодо дитини.
4. Великі компанії технічного характеру – це можливість API-інтеграції, з подальшим монтуванням прогнозувальних алгоритмів у вже існуючі великі системи.
5. Державні установи – це можливість партнерств у великих національних програмах.

Усі вищезазначені пункти мають можливість розширення користувацької бази та аудиторії споживачів, а також їх довіри до ПП.

### **6.2.9. Витрати**

За часовий проміжок для створення, випуск у світ та подальшу роботу стартап-проєкту було взято дев'ять місяців. З урахуванням цього, маємо наступні витрати:

1. Розробка системи прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників та технічне впровадження, в

тому числі MVP веб-версії платформи, мобільного додатку, додаткових модулів та інтеграційних можливостей – від 480 000 до 720 000 гривень.

2. Оплата праці команди терміном в дев'ять місяців, у тому числі керівника проєкту, аналітика, full-stack розробника, інженера у сфері машинного навчання, дизайнера, фахівця з безпеки інформації, консультантів та технічного фахівця – від 800 000 до 1 000 000 гривень.

3. Маркетинг та стратегії поширення, у тому числі соціальні мережі, реклама, власний бренд – від 200 000 до 320 000 гривень.

4. Організація та юридичні питання – від 80 000 до 120 000 гривень.

5. Технічне забезпечення – від 80 000 до 160 000 гривень.

6. Резервний фонд проєкту – приблизно 120 000 гривень.

Підсумовуючи, отримуємо загальну орієнтовну суму витрат – від 1 760 000 до 2 440 000 гривень.

#### **6.2.10. Споживчі властивості товару**

Отримана система прогнозування проблемного рівня використання Інтернету дітьми – це унікальний продукт, наразі єдиний відомий в ниші за тематикою.

Ключовими унікальними властивостями ПП є:

1. Використання унікальної комбінації параметрів (у тому числі фізичних), в якості індикаторів можливої Інтернет-залежності, в поєднанні з методами машинного навчання.

2. Безпека даних, відсутність порушення цифрової конфіденційності дитини, не потрібне відстеження споживаного контенту.

3. Персоналізований підхід – результат прогнозування залежить від індивідуальних показників кожного окремого користувача, надання рекомендацій щодо подальших дій.

4. Профілактичні дії – виявлення початкових маркерів виникнення залежності від мережі Інтернет, можливість раннього втручання.

5. Зручність інтерфейсу та системи загалом, адаптовність під різні пристрої.

6. Подальша можливість інтеграції з носимими пристроями – неперервний збір та аналіз даних без додаткового втручання.

#### **6.2.11. Дослідження ринку**

Розглядаючи поточний ринок цифрових розробок у сфері виявлення залежності від мережі Інтернет та різного роду розладів, переважають рішення, що ґрунтуються на моніторингу поведінки в Інтернеті, а саме: екранний час, контент, який споживає дитина, інша активність в мережі тощо. Такий підхід жодним чином не враховує фізіологічних змін у тілі дитини, які можуть стати ключовими індикаторами у поставленій задачі передбачення. Цю нішу й займає розроблюваний ПП, який стане першим, унікальним та інноваційним рішенням в даній сфері.

За конкурентів можна вважати такі сервіси, як Bark, Google Family Link, Qustodio, що пропонують функціонал батьківського контролю, але, як уже було зазначено, не враховують величезний спектр ознак – лише часткова об'єктивність. Натомість, в сучасній медицині широко застосовуються психологічні опитувальники та анкети, які також не можна вважати об'єктивними, так як усе залежить від власного сприйняття себе учасником.

Підсумовуючи, конкурентних рішень, які б надавали подібний до нашого стартап-проєкту функціонал, наразі виявлено не було.

Оскільки цей ПП є інноваційним, дуже важливу роль буде грати маркетингова стратегія його просування, до якої входитимуть наступні види діяльності:

1. Пілотні програми;
2. Реклама для цільової аудиторії;
3. Інформаційні партнерства;
4. Профільні заходи;
5. Реферальні програми;

## 6. Сертифікація продукту.

### 6.2.12. Елементи фінансового плану

#### 1. Опис бізнес-проєкту.

Передбачається реалізація цифрової системи, що надаватиме функціонал з прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників. Проєкт у своїй основі є технологічним стартапом, з орієнтацією на сегменти B2C і B2B. Мета – ПП для раннього виявлення залежності від мережі Інтернет, основними користувачами якого можуть стати батьки, вчителі та педагоги, психологи, медичні працівники.

#### 2. Опис продукту.

Продукт – це аналітична платформа, доступ до якої виконуватиметься через вебсайт, десктопний застосунок чи мобільний додаток. В основі лежать підходи до машинного навчання та роботи з різними видами даних. Передбачення інтерпретуються, створюються рекомендації щодо подальших дій, можливостей звернення до спеціалістів. Додаткові модулі під певні категорії споживачів.

#### 3. Маркетинг та продажі.

Дослідження наявних підходів і рішень щодо виявлення та профілактики Інтернет-залежності, показало різкий стрибок попиту на подібні рішення, а також важливість збереження цифрової приватності дитини. Проблема, що вирішується, наразі є глобальною, проте аналіз конкурентів показав, що всі вони спираються на поведінковий моніторинг. Дана система є інноваційною та першою, як в Україні, так і в усьому світі (існують лише експериментальні дослідження та проєкти).

#### 4. Фінансовий план.

За попередніми розрахунками орієнтовна загальна сума витрат – від 1 760 000 до 2 440 000 гривень.

#### 5. Термін окупності проєкту.

Оцінюючи актуальність роботи та попит на подібні рішення, різке зростання масштабів проблеми, яку вирішує наша система, приблизний термін окупності проєкту становить від 12 до 18 місяців з моменту старту. Цей план досягається за допомогою гнучкості моделі монетизації, активним маркетинговим стратегіям та значущості проєкту для населення. Також, велика кількість інвестицій та партнерств може пришвидшити отримання плюсових результатів.

#### 6. Можливі ризики та методи їх вирішення.

Ризик початкової недостатності попиту – пілотні впровадження, маркетингові стратегії, низька ймовірність через цінність продукту.

Ризик технічних помилок – наявність висококваліфікованих фахівців, завчасне тестування продукту, поетапне впровадження, резервне копіювання матеріалів та вихідних кодів, постійні покращення та оновлення.

Ризик недовіри до приватності даних – політика конфіденційності інформації, відповідність стандартам.

Ризик конкуренції – унікальність підходу, відсутність плагіату, швидкі покращення після отримання зворотного зв'язку, важливі партнерства.

Ризик недостатності фінансування – пошук нових джерел отримання коштів, слідування плану просування продукту, готовність до ризиків та додаткові стратегії на такі випадки.

#### 6.2.13. Резюме

Стартап-проєкт на тему «Система прогнозування рівня проблемного використання Інтернету дітьми на основі фізичних показників» має на меті створення інноваційного ПП для батьків, вчителів та педагогів, психологів та медичних працівників, що матиме функціонал для визначення рівня проблемного використання мережі та раннього виявлення розладів, пов'язаних із Інтернет-залежністю. Унікальність підходу до прогнозування базується саме на використанні фізичних показників як ключових індикаторів для визначення ризик-профілю кожної окремої дитини.

Втручання у приватне цифрове життя дитини, контроль споживаного контенту, батьківський контроль – усе це стає необов'язковим за допомогою нашої системи, використовується лише об'єктивний, науково підтверджений підхід. Висока довіра з боку користувачів та дітей – є одним з ключових аспектів системи.

Дослідження ринку конкурентів показало – ПП являтиметься першим в своєму роді та сфері. Жодних конкурентних рішень не було виявлено як в Україні, так і в усьому світі, лише частково функціональні аналоги. Актуальність та висока соціальна важливість проблеми, гнучкість монетизаційної моделі, маркетингові стратегії та інших чинників роблять цей проєкт перспективним як з точки зору потенціалу просування в соціумі, так і зі сторони комерції.

Інвестиції у цей стартап-проєкт відкривають не тільки фінансові перспективи, але можуть стати справжнім внеском до скарбниці покращення життя населення, а також кроком до формування всебічно здорового молодого покоління в епоху правління цифрових технологій.

## **Висновки до розділу 6**

У цьому розділі було описано стартап-проєкт на основі результатів МД. Створено бізнес-модель майбутнього проєкту, вказано цінність продукту, ключові ресурси та партнерів, канали збуту та сегменти користувачів.

Розрахунки економічної частини довели раціональність реалізації та потенційну окупність проєкту. Викладена система прогнозування рівня проблемного використання Інтернету дітьми має потенціал реалізації як в якості самостійного продукту, так і модуля для подальшої інтеграції у більші медико-психологічні та освітні системи.

Стартап-проєкт має високу соціальну значущість суспільства, може стати унікальним рішенням в сфері цифрової безпеки та сприяти формуванню культури раціонального користування сучасних технологій серед молоді.

## ЗАГАЛЬНІ ВИСНОВКИ

В результаті проведеного дослідження можна беззаперечно стверджувати потенціал використання методів машинного навчання в поєднанні з різноманітними видами даних, зокрема, фізичними, у контексті задачі прогнозування рівня проблемного використання Інтернету дітьми. Робота поєднує в собі огляд та аналіз вже існуючих підходів та засобів розв'язання поставленої задачі, а також практичну реалізацію програмного застосунку для прогнозування рівня проблемного використання мережі Інтернет.

Аналітичний огляд повноцінно обґрунтував актуальність тематики дослідження, спираючись на світові тенденції розвитку сучасних, цифрових технологій. Статистичні дані вказують на постійний ріст кількості дітей та підлітків, залежних від Інтернету, тим самим закріплюючи початкові припущення. З іншої сторони – наукові дослідження з використанням технологій машинного навчання, що на своєму прикладі доводять ефективність застосування таких підходів в комбінації з фізичними, психологічними та іншими даними.

Методи, алгоритми та підходи машинного навчання – ті самі, які використовують нинішні науковці та дослідники, були детально розглянуті в якості засобів реалізації алгоритму прогнозування. Було описано базові алгоритми, удосконалені поодинокі моделі, ансамблеві підходи, нейронні мережі, а також схему проведення експериментів та валідації – все необхідне для подальшого практичного застосування.

Далі було обґрунтовано вибір мови програмування – Python; бібліотек – NumPy, Pandas, Tabulate, Scikit-learn, PyTorch, Skorch, XGBoost, LightGBM, CatBoost, Matplotlib, Seaborn, Tkinter, Joblib; середовища розробки – PyCharm.

За набір даних для роботи було вирішено взяти датасет з інформацією щодо проблемного використання Інтернету дітьми, зі змагання від Child Mind Institute. Дані було проаналізовано, побудовано відповідні графіки та

статистики, оброблено за використання сучасних підходів, заповнено пропущені значення, видалено надлишкові та шкідливі змінні, зібрано у навчальну та тестову вибірки.

Наступним етапом була робота з різноманітними алгоритмами машинного навчання (а також створено спеціалізовану схему для оцінки якості їх роботи) – базові моделі класифікації та регресії, їх оптимізація та підбір гіперпараметрів, створення ансамблів декількох типів, використовуючи різні набори алгоритмів та параметрів. Також було розглянуто роботу нейронних мереж класифікації, регресії, та їх ансамблю – від базових, до більш складних архітектур.

В результаті, для подальшої роботи, було обрано регресійну модель градієнтного бустингу CatBoost, з підібраними гіперпараметрами, протестовано на відповідній вибірці, що показала помірний рівень узгодженості своїх прогнозів з реальними даними за коефіцієнтом згоди Каппа (0.426). Фінальну модель та все необхідне було збережено в файл для подальшого використання.

Наступним кроком стало створення програмного застосунку, який, використовуючи кінцеву модель, прогнозує рівень проблемного використання Інтернету дитиною. Додаток містить зручний та зрозумілий візуальний інтерфейс, а також наступний функціонал: зчитування даних з CSV-файлу, або ручне введення, їх валідація та обробка, надання до моделі машинного навчання, та, в результаті, виведення прогнозу та рекомендацій.

Далі, було узагальнено результати, отримані на всіх етапах проведення дослідження, проведено їх аналіз, наведено унікальність цієї роботи, а саме: схеми роботи з даними, конфігурації моделі, підвищення точності результатів роботи моделі CatBoost за своїх попередників на 4.25 відсотки. Описано практичне значення результатів та їх можливе застосування в майбутньому.

Останнім етапом стало створення та опис стартап-проєкту за темою роботи.

Підсумовуючи, дане дослідження повністю підтвердило актуальність тематики прогнозування індикаторів проблемного використання Інтернету у дітей, а також використання методів машинного навчання в поєднанні з фізичними даними в контексті поставленої задачі. Таким чином, було створено повноцінно функціонуючий ПЗ, який в подальшому може використовуватися як окремо, так і інтегруючись у різноманітні системи, сприяючи підвищенню благополуччя молодого покоління.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Young, K. (1998). Internet Addiction: The Emergence of a New Clinical Disorder. *CyberPsychology & Behavior*, 1(3), 237–244.  
URL: <https://www.liebertpub.com/doi/10.1089/cpb.1998.1.237> (дата звернення 01.09.2025)
2. Welch, V., Wy, T. J., Ligezka, A., Hassett, L. C., Croarkin, P. E., Athreya, A. P., & Romanowicz, M. (2022). Use of Mobile and Wearable Artificial Intelligence in Child and Adolescent Psychiatry: Scoping Review. *Journal of Medical Internet Research*, 24(3), e33560.  
URL: <https://www.jmir.org/2022/3/e33560> (дата звернення 01.09.2025)
3. Gu, H., Shi, B., He, H., Yuan, S., Cai, J., Chen, X., & Wan, Z. (2025). Association Between Excessive Internet Use Time, Internet Addiction, and Physical-Mental Multimorbidity Among Chinese Adolescents: Cross-Sectional Study. *Journal of Medical Internet Research*, 27, e69210.  
URL: <https://www.jmir.org/2025/1/e69210> (дата звернення 01.09.2025)
4. Lozano-Blasco, R., Latorre-Martínez, M., & Cortés-Pascual, A. (2022). Screen addicts: A meta-analysis of internet addiction in adolescence. *Children and Youth Services Review*, 135, 106373.  
URL: <https://www.sciencedirect.com/science/article/pii/S0190740922000093> (дата звернення 02.09.2025)
5. Gan, Y., Kuang, L., Xu, X.-M., Ai, M., He, J.-L., Wang, W., Hong, S., Chen, J. mei, Cao, J., & Zhang, Q. (2025). Application of machine learning in predicting adolescent Internet behavioral addiction. *Frontiers in Psychiatry*, 16, 1521051.  
URL: <https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsy.2024.1521051/full> (дата звернення 03.09.2025)
6. Wikipedia Contributors. (2019, February 8). Internet addiction disorder. Wikipedia; Wikimedia Foundation.

- URL: [https://en.wikipedia.org/wiki/Internet\\_addiction\\_disorder](https://en.wikipedia.org/wiki/Internet_addiction_disorder) (дата звернення 03.09.2025)
7. World Health Organization. (2023). Gaming Disorder. World Health Organization.  
URL: <https://www.who.int/standards/classifications/frequently-asked-questions/gaming-disorder> (дата звернення 03.09.2025)
  8. Greenwood, S. (2022, December 6). Internet, smartphone and social media use. Pew Research Center's Global Attitudes Project.  
URL: <https://www.pewresearch.org/global/2022/12/06/internet-smartphone-and-social-media-use-in-advanced-economies-2022> (дата звернення 03.09.2025)
  9. Українська правда. Життя. (2024, September 25). Кожен десятий підліток має ознаки залежності від соцмереж чи ігор в інтернеті: звіт ВООЗ. Українська правда. Життя.  
URL: <https://life.pravda.com.ua/society/skilki-pidlitkiv-mayut-zalezhnist-vid-internetu-ta-onlayn-igor-zvit-vooz-303934/> (дата звернення 03.09.2025)
  10. Chen, S., Weng, L., Su, Y., Wu, H., & Yang, P. (2003). Development of a Chinese Internet Addiction Scale and Its Psychometric Study. Chinese Journal of Psychology, 45, 279-294.  
URL: [https://www.researchgate.net/publication/271194311\\_Development\\_of\\_Chinese\\_Internet\\_Addiction\\_Scale\\_and\\_its\\_psychometric\\_study](https://www.researchgate.net/publication/271194311_Development_of_Chinese_Internet_Addiction_Scale_and_its_psychometric_study) (дата звернення 04.09.2025)
  11. Caplan, S. E. (2010). Theory and measurement of generalized problematic Internet use: A two-step approach. Computers in Human Behavior, 26(5), 1089–1097.  
URL: <https://www.sciencedirect.com/science/article/abs/pii/S074756321000052X> (дата звернення 04.09.2025)

12. Demetrovics, Z., Szeredi, B., & Rózsa, S. (2008). The three-factor model of Internet addiction: The development of the Problematic Internet Use Questionnaire. *Behavior Research Methods*, 40(2), 563–574. URL: <https://link.springer.com/article/10.3758/BRM.40.2.563> (дата звернення 04.09.2025)
13. Razavi, M., McDonald, A., Mehta, R., & Sasangohar, F. (2023, September 20). Evaluating Mental Stress Among College Students Using Heart Rate and Hand Acceleration Data Collected from Wearable Sensors. *ArXiv.org*. URL: <https://arxiv.org/abs/2309.11097> (дата звернення 04.09.2025)
14. Xiao, B., Parent, N., Rahal, L., & Shapka, J. (2023). Using Machine Learning to Explore the Risk Factors of Problematic Smartphone Use among Canadian Adolescents during COVID-19: The Important Role of Fear of Missing Out (FoMO). *Applied Sciences*, 13(8), 4970–4970. URL: <https://www.mdpi.com/2076-3417/13/8/4970> (дата звернення 04.09.2025)
15. Wong, E. (2024, December 12). Predicting Problematic Internet Use in Children. *Medium*. URL: <https://medium.com/@ethanrwong/predicting-problematic-internet-use-in-children-ffce468cd155> (дата звернення 05.09.2025)
16. Wang, W., Chen, J., Hu, Y., Liu, H., Chen, J., Thippa Reddy Gadekallu, Garg, L., Mohsen Guizani, & Hu, X. (2024). Integration of Artificial Intelligence and wearable Internet of Things for mental health detection. *International Journal of Cognitive Computing in Engineering*. URL: <https://www.sciencedirect.com/science/article/pii/S2666307424000238> (дата звернення 05.09.2025)
17. Ahmed, A., Agus, M., Alzubaidi, M., Aziz, S., Abd-Alrazaq, A., Giannicchi, A., & Househ, M. (2022). Overview of the role of big data in mental health: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2, 100076.

- URL: <https://www.sciencedirect.com/science/article/pii/S2666990022000271>  
(дата звернення 05.09.2025)
18. Nawi, A., Hussin, Z., Ren, C. C., Norsaidi, N. S., & Mohd Pozi, M. S. (2020). Identifying the Types of Digital Footprint Data Used to Predict Psychographic and Human Behaviour. *Digital Libraries at Times of Massive Societal Transition*, 287–296.  
URL: [https://dl.acm.org/doi/10.1007/978-3-030-64452-9\\_26](https://dl.acm.org/doi/10.1007/978-3-030-64452-9_26) (дата звернення 05.09.2025)
  19. Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., & Chan, E. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4(1), 170181.  
URL: <https://www.nature.com/articles/sdata2017181> (дата звернення 05.09.2025)
  20. GeeksforGeeks. (2018, September 13). Linear Regression in Machine learning. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/ml-linear-regression/>  
(дата звернення 08.09.2025)
  21. GeeksforGeeks. (2017, May 9). Logistic Regression in Machine Learning. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/> (дата звернення 08.09.2025)
  22. GeeksforGeeks. (2017, October 16). Decision Tree. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/decision-tree/> (дата звернення 08.09.2025)
  23. GeeksforGeeks. (2017, April 14). KNearest Neighbor(KNN) Algorithm. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/>  
(дата звернення 08.09.2025)

24. GeeksforGeeks. (2021, January 20). Support Vector Machine (SVM) Algorithm. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/> (дата звернення 08.09.2025)
25. GeeksforGeeks. (2025, January 13). Types of Ensemble Learning. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/types-of-ensemble-learning/> (дата звернення 10.09.2025)
26. GeeksforGeeks. (2024, February 22). Random Forest Algorithm in Machine Learning. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/> (дата звернення 10.09.2025)
27. GeeksforGeeks. (2020, August 25). Gradient Boosting in ML. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting/> (дата звернення 10.09.2025)
28. GeeksforGeeks. (2021, September 18). XGBoost. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/xgboost/> (дата звернення 10.09.2025)
29. GeeksforGeeks. (2020, July 15). LightGBM (Light Gradient Boosting Machine). GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/lightgbm-light-gradient-boosting-machine/> (дата звернення 10.09.2025)
30. GeeksforGeeks. (2021, January 20). CatBoost in Machine Learning. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/catboost-ml/> (дата звернення 10.09.2025)
31. GeeksforGeeks. (2019, January 17). What is a Neural Network? GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/machine-learning/neural-networks-a-beginners-guide/> (дата звернення 11.09.2025)

32. GeeksforGeeks. (2023, November 29). Training data vs Testing data. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/python/training-data-vs-testing-data/> (дата звернення 11.09.2025)
33. GeeksforGeeks. (2024, March 19). How to Calculate Cohen's Kappa in R. GeeksforGeeks.  
URL: <https://www.geeksforgeeks.org/r-language/how-to-calculate-cohens-kappa-in-r/> (дата звернення 11.09.2025)
34. Comparison of Programming Languages. (2023).  
URL: <https://www.amorserv.com/insights/comparison-of-programming-languages> (дата звернення 12.09.2025)
35. The Python Standard Library. (2025). Python Documentation.  
URL: <https://docs.python.org/3/library> (дата звернення 12.09.2025)
36. Py-Core Python Programming. (2025, January 9). Python and IDEs – A Rundown of Development Environment Options. Medium.  
URL: <https://medium.com/@ccpythonprogramming/python-and-ides-a-rundown-of-development-environment-options-10e62b6e0b72> (дата звернення 12.09.2025)
37. Child Mind Institute – Problematic Internet Use. (2014). Kaggle.  
URL: <https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use> (дата звернення 16.09.2025)
38. Kitazawa, M., Yoshimura, M., Murata, M., Sato-Fujimoto, Y., Hitokoto, H., Mimura, M., Tsubota, K., & Kishimoto, T. (2018). Associations between problematic Internet use and psychiatric symptoms among university students in Japan. *Psychiatry and Clinical Neurosciences*, 72(7), 531–539.  
URL: <https://onlinelibrary.wiley.com/doi/10.1111/pcn.12662> (дата звернення 16.09.2025)
39. GeeksforGeeks. (2020, July 31). Missing data imputation with fancyimpute. GeeksforGeeks.

URL: <https://www.geeksforgeeks.org/machine-learning/missing-data-imputation-with-fancyimpute/> (дата звернення 19.09.2025)

40. GeeksforGeeks. (2024, June 28). Handling Missing Data with IterativeImputer in Scikitlearn. GeeksforGeeks.

URL: <https://www.geeksforgeeks.org/machine-learning/handling-missing-data-with-iterativeimputer-in-scikit-learn/> (дата звернення 19.09.2025)

41. GeeksforGeeks. (2020, July 15). StandardScaler, MinMaxScaler and RobustScaler techniques ML. GeeksforGeeks.

URL: <https://www.geeksforgeeks.org/machine-learning/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/> (дата звернення 27.09.2025)