Вінницький національний технічний університет
Факультет комп'ютерних систем і автоматики
Кафедра системного аналізу та інформаційних технологій

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему:

«Інформаційна технологія аналізу рівня знань іноземної мови студентів закладу вищої освіти»

Виконав: студент 2 курсу, групи 2ІСТ-20м
спеціальності 126 – «Інформаційні системи та технології»
_____ Лотоцький А. О.

Керівник: к.т.н., доц. каф. САІТ
_____Козачко О. М.
«___» _____ 2021 р.

Опонент: к.т.н., доц. каф. АІІТ
_____ Кабачій В. В.
«___» _____ 2021 р.

Допущено до захисту

Завідувач кафедри САІТ
_____д.т.н., проф. Мокін В. Б.
«___» _____ 2021 р.

Вінниця ВНТУ – 2021 рік

Вінницький національний технічний університет
Факультет комп'ютерних систем і автоматики
Кафедра системного аналізу та інформаційних технологій
Рівень вищої освіти – ІІ-й (магістерський)
Галузь знань – 12 Інформаційні технології
Спеціальність – 126 Інформаційні системи та технології
Освітньо-професійна програма – Інформаційні технології аналізу даних та зображень

ЗАТВЕРДЖУЮ
Завідувач кафедри САІТ
_____д.т.н., проф. Мокін В. Б.
«___» _____ 2021 р.

ЗАВДАННЯ
НА МАГІСТЕРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
Лотоцькому Андрію Олександровичу

1. Тема роботи: «Інформаційна технологія аналізу рівня знань іноземної мови студентів закладу вищої освіти»,
керівник роботи: Козачко О. М., к.т.н., доц. каф. САІТ,
затверджені наказом закладу вищої освіти від «___»_____2021 року №__
2. Строк подання студентом роботи «___»_____2021 року
3. Вихідні дані до роботи:

Датасети з даними про оцінювання студентів Вінницького національного медичного університету ім. М. І. Пирогова за 2, 4, 6 семестри навчання та результати проходження першого етапу ЄДКІ з англійської мови.

4. Зміст текстової частини:
- аналіз предметної області;
- розробка методу дослідження даних;
- розробка моделей виявлення закономірностей;
- реалізація програмного модуля аналізу рівня знань іноземної мови студентів закладу вищої освіти.
- економічна частина.

5. Перелік ілюстративного матеріалу (з точним зазначенням обов'язкових креслень):
- діаграма важливості ознак lightgbm моделі;
- діаграма важливості ознак xgboost моделі;
- матриця невідповідності логістичної регресії;
- матриця невідповідності random forest;
- дерево рішень;
- реляційна структура бази даних;
- загальний вигляд інтерфейсу веб-ресурсу.

6. Консультанти розділів МКР

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
5	Ратушняк О.Г., к.е.н., доц. каф. ЕПВМ		

7. Дата видачі завдання «___»_____2021 року

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів МКР	Строк виконання етапів роботи	Примітка
1	Аналіз предметної області	09.2021	
2	Розробка методу дослідження даних	09.2021	
3	Розробка моделей виявлення закономірностей	09.2021	
4	Реалізація програмного модуля аналізу рівня знань іноземної мови студентів закладів вищої освіти	10.2021	
5	Економічна частина	10.2021	
6	Оформлення матеріалів до захисту МКР	11.2021	

Студент _____ Лотоцький А. О.

Керівник роботи _____ Козачко О. М.

АНОТАЦІЯ

УДК 004.08

Лотоцький А. О. Інформаційна технологія аналізу рівня знань іноземної мови студентів закладу вищої освіти. Магістерська кваліфікаційна робота зі спеціальності 126 – інформаційні системи та технології, освітньо-професійна програма – інформаційні технології аналізу даних та зображень. Вінниця: ВНТУ, 2021. 108 с.

На укр. мові. Бібліогр.: 21 назв; рис.: 109; табл.: 5.

В магістерській кваліфікаційній роботі виконано аналіз предметної галузі аналізу рівня знань та обґрунтовано актуальність даного дослідження. Проведено розвідувальний аналіз даних та відібрано оптимальний набір ознак, за якими слід будувати модель. Проведено аналіз існуючих методів аналізу для розв’язання поставленої задачі. Розроблено та наведено комп’ютерну програму, яка розв’язує поставлену задачу оптимальним методом. Об’єкт досліджень є процес визначення факторів, що впливають на оцінювання іноземної мови студентів Вінницького національного медичного університету ім. М. І. Пирогова за 2, 4, 6 семестри навчання та результати першого етапу ЄДКІ.

Ілюстративна частина складається з 7 плакатів із результатами моделювання.

У розділі економічної частини розглянуто питання про доцільність розробки та впровадження інформаційна технологія аналізу рівня знань іноземної мови студентів закладу вищої освіти.

Ключові слова: аналіз даних, розвідувальний аналіз, побудова моделей, виявлення закономірностей, щодо факторів впливу, розробка програмного модуля

ABSTRACT

Lototskyi A. O. Information technology of analysis of the level of foreign language knowledge of students of higher education institutions.

Master's qualification work in the specialty 126 – information systems and technologies, educational and professional program – information technologies of data and image analysis. Vinnytsia: VNTU, 2021. – 108 p.

In Ukrainian language. Bibliographer: 21 titles; fig.: 109; table: 5.

In the master's qualification work the analysis of the subject area of the analysis of the level of knowledge is performed and the relevance of this research is substantiated. An exploratory analysis of the data was conducted and the optimal set of features on which to build the model was selected. Performed an analysis of existing methods of analysis to solve the problem. A computer program has been developed and presented, which solves the problem in an optimal way. The object of research is the process of determining the factors influencing the assessment of foreign languages of students of Vinnytsia National Medical University. MI Pirogov for 2, 4, 6 semesters of study and the results of the first stage of EDKI.

The illustrative part consists of 7 posters with simulation results.

The economic part considers the feasibility of developing and implementing an information technology of analysis of the level of foreign language knowledge of students of higher education institutions.

Key words: data analysis, exploratory data analysis, model building, identification of patterns of influencing factors, software module development.

ЗМІСТ

ВСТУП

Актуальність теми. З стрімким розповсюдженням інформаційних технологій, збільшується і кількість збереженої інформації. Ця інформація може бути використана для проведення аналізу, тобто, виявлення загальних тенденцій змін, факторів, що впливають на них. За допомогою виявлених тенденцій та факторів можна оптимізувати роботу системи, підвищити продуктивність, скоротити кількість втрат. Також, сфера дослідження є однією з найбільш перспективних. Для будь-кого є очевидною важливість вивчення англійської мови в наш час. Англійська мова відкриває нескінченні можливості у повсякденному та професіональному житті, тому проблема є актуальною.

Об'єктом дослідження є процес аналізу рівня знань іноземної мови студентів Вінницького національного медичного університету ім. М. І. Пирогова за 2, 4, 6 семестри навчання та результати першого етапу ЄДКІ.

Предмет дослідження – інформаційні технології аналізу рівня знань іноземної мови студентів закладу вищої освіти.

Метою роботи є виявлення закономірностей щодо факторів, які впливають на підвищення рівня знань іноземної мови.

Завданнями роботи відповідно до поставленої мети є різносторонній аналіз даних: здійснення аналізу предметної галузі, розвідувального аналізу, побудова моделі об'єкта та визначення факторів, що впливають на рівень вивчення іноземної мови. На основі цих досліджень можна буде зробити адекватні і достатньо точні висновки про виявлені закономірності.

Новизна одержаних результатів. Запропоновано моделі виявлення закономірностей факторів, що впливають на рівень знань іноземної мови студентами закладів вищої освіти, які на відмінну від існуючих, використовують методи машинного навчання та забезпечують виявлення дисциплін, які найбільше впливають на рівень знань іноземної мови.

Практичне значення роботи полягає у можливості використання розробленої інформаційної технології, для визначення навчальних предметів, що найбільше впливають на результати оцінювання з англійської мови.

Апробація результатів магістерської кваліфікаційної роботи. Результати кваліфікаційної роботи доповідались на Всеукраїнській науково-практичній інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи» (Вінниця, 2021-2022 рр.).

Публікації результатів магістерської кваліфікаційної роботи. Опубліковано тези в збірнику матеріалів Всеукраїнської науково-практичної інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи» (Вінниця, 2021-2022 рр.) [1].

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Огляд предметної області аналізу

Враховуючи виклики та проблеми, з якими стикається сучасний навчальний процес, підвищується рівень використання сучасних інтелектуальних систем та алгоритмів для підвищення рівня освіти та викладання у навчальних закладах. Існує велика кількість досліджень направлених на визначення закономірностей у цій сфері. Ці дослідження можуть бути ефективно використані для з'ясування та виявлення сучасних освітніх проблем, а також індивідуальних та колективних особливостей учнів та студентів, за допомогою впровадження процесу класифікації та регресійного аналізу набору даних.

Розглянемо кілька проблем та підходів до їх вирішення, які допоможуть визначити напрям ведення дослідження та запобігти виникнення схожих проблем.

Першим прикладом вирішення схожої задачі постає наукова стаття під назвою «Predicting Pupil’s Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods». У роботі описано та продемонстровано результати, отримані від використання алгоритмів, аналізу даних. Визначено основні особливості, залежності та методи виділення основних ознак та факторів із набору даних. Прогнозування характеристик учнів допоможе поділити їх на різні класи, щоб це дозволило учням розвивати комунікативні, лідерські та навички самоврядування під час навчання в школі чи університеті. Результати показують, що оцінювання показників ефективності є невід’ємною частиною покращення сучасного навчального процесу [2].

Наступним прикладом є наукова стаття «Mathematics Learning through Computational Thinking Activities», у якій, в свою чергу, розглянуто та систематизовано багато прикладів та ідей з інших джерел. Обчислювальне

мислення являє собою термінологію, яка охоплює складний набір процесів міркувань, які проводяться для постановки проблеми та вирішення за допомогою обчислювального інструменту. Здатність систематизувати задачі та розв’язувати їх цими засобами наразі вважається навиком, який повинен розвивати всі студенти разом із мовою, математикою та науками. Враховуючи, що інформатика має багато коренів у галузі математики, розумно розмірковувати, чи можна впливати на навчання математики, пропонуючи студентам заходи, пов'язані з обчислювальним мисленням. У цьому сенсі у цій статті представлено систематичний огляд літератури щодо повідомлених доказів навчання математики в діяльності, спрямованій на розвиток навичок обчислювального мислення. Проаналізовано сорок дві статті, в яких були представлені рішення для оцінки результатів навчання, опубліковані з 2006 по 2017 рік [3].

У статті «Teaching computational thinking to english learners» розглянуто, що є спільного у розвинені комп’ютерного мислення та вивченні англійської мови, які методи є найбільш ефективними та простими. Ефективне навчання обчислювального мислення для учнів англійської мови суттєво перегукується з іншими формами контентного навчання. Аналіз комп'ютерного коду можна використовувати для побудови мета інформованості обчислювальної семіотики, а візуальна природа деяких мов програмування, таких як Scratch, може сприяти розвитку грамотності. Найголовніше, що проекти, пов'язані з обчислювальним мисленням, чи то в створенні розповідей, чи в розробці електронних текстових проектів, дають студентам широкі можливості висловити та розвинути власну ідентичність - важливий елемент успішної програми з другої мови. Зараз формується все поле обчислювального мислення в освіті [4].

Цікаві результати винесено у статті «An overview of Machine Learning Technologies and their use in E-learning». В ній проведено огляд основних алгоритмів машинного навчання для загальної класифікації студентів. Найкращий результат, з точністю 97%, показали метод к-середніх та наївний

баєсів класифікатор. Нижчу точність показали класифікатор градієнтного бустингу з точністю 82%, та метод k-найближчих сусідів – 85% [5].

Схожий аналіз проведено в статті «Predicting academic performance using automatic learning techniques: a review of the scientific literature». В ньому проведено систематичний огляд літератури з використання методів машинного навчання для прогнозування успішності. Було виявлено що найбільш використовуваними є лінійна регресія, наївний баєсів класифікатор та randomized forest. Також було показано, що зазвичай для вирішення поставленої задачі використовується більше ніж один метод машинного навчання [6].

Також існують вітчизняні дослідження в цій сфері. Першим прикладом є стаття під назвою «Використання штучних нейронних мереж в оцінці показників діяльності закладів вищої освіти». В ній розглянуто використання методів та засобів машинного навчання, інтелектуального аналізу, які використовують інструментарій штучних нейронних мереж для формування, розпізнавання та аналізу образів різних явищ. Зокрема використання нейронних мереж для прогнозування та аналізу прогнозування результатів діяльності в галузі освіти. Також розглянуто можливі методи застосування штучного інтелекту, такі як: прогнозування успішності студентів, створення чат-ботів, профілювання та групування, надання звітів, прогнозування інтересів студентів. Визначено бібліотеки мови Python, що спрощують аналіз та моделювання. Наприклад Pandas та Numpy, перша з них дозволяє прискорити завантаження даних їх обробку для розвідувального аналізу, а друга використовується для роботи з таблицями, числовими даними та автоматизації типових операцій зв’язаних їх обробкою [7].

Наступною розглянутою статтею є «Нейромережеве прогнозування складання студентами медиками ліцензійного інтегрованого іспиту «крок 1» на основі результатів поточної успішності та семестрового комплексного тестового іспиту». В ній використовується доволі застаріла технологія NeuroXL – це нейромережевий пакет для Microsoft Excel. В роботі розглянуто

його можливості для прогнозування ліцензійного інтегрованого іспиту на основі успішності студентів та результатів семестрових іспитів. Недоліком статті є суттєвє розходження результатів прогнозування з реальними результатами для деяких студентів [8].

У статті «Прогнозування успішності навчання студентів – один із напрямів підвищення якості освіти» розглянуто та доведено, що якість проведення навчальних занять має прямий вплив на успішність навчання студентів. За допомогою кореляційного аналізу було встановлено, що якість проведення лабораторно-практичних занять має найбільший кореляційний зв’язок з успішністю. А проведення лекційних занять, в свою чергу, має менший вплив [9].

1.2 Огляд і опис методу прогнозування

Машинне навчання – це клас методів штучного інтелекту, характерною рисою яких є не пряме рішення задачі, а навчання в процесі застосування рішень безлічі подібних завдань. Для побудови таких методів використовуються засоби математичної статистики, чисельних методів, методів оптимізації, теорії ймовірностей, теорії графів, різні техніки роботи з даними в цифровій формі [10].

Фахівець з Machine learning повинен збирати, систематизувати та аналізувати дані, а потім на основі отриманої інформації створювати алгоритми штучного інтелекту. Машинне навчання стало проривом в аналітиці: завдяки йому нарешті виявилось працювати з неструктурованими даними - зображеннями, звуками тощо. Машинне навчання знаходиться на стику математики, програмування, штучного інтелекту і ніколи не було ізольованим від вирішуваної проблеми. Основними тенденціями машинного навчання є нейронні мережі. Вони працюють з неструктурованою інформацією, такою як зображення та текстові дані, подібно до людського

мозку. Це дає розуміння того, як працює глибоке навчання, сприяє його широкому розвитку та використанню. Вони допомагають у виборі архітектури та параметрів. Нейронні мережі також можуть бути закодовані в Python.

Основними задачами машинного навчання можна вважати:

– Задачу класифікації;
– Задачу відновлення регресії;
– Задачу кластеризації;
– Задачу ідентифікації;
– Задачу прогнозування.

Задача прогнозування, яка нас і цікавить, історично виникла при дослідженні часових рядів і спробі передбачення їх значень через якийсь проміжок часу [11].

Існують різні методи побудови моделі подібних об’єктів, у тому числі з використанням методі машинного навчання, наприклад:

– Нейронні мережі: обчислення структуруються в термінах взаємозв'язаних груп штучних нейронів, які обробляють інформацію із застосуванням конективістського підходу до обчислень. Сучасні нейронні мережі є нелінійними статистичними інструментами моделювання даних. Їх зазвичай застосовують для моделювання складних взаємозв'язків між входами та виходами, для пошуку закономірностей в даних, або для виявлення статистичної структури в невідомому спільному розподілі ймовірності спостережуваних величин;

– Дерева рішень: використовує як передбачувальну модель дерево рішень, яке відображує спостереження про предмет на висновки про цільове значення предмету;

– Баєсові мережі: ймовірнісна графова модель, яка представляє набір випадкових величин та їхніх умовних незалежностей через спрямований ациклічний граф.

Оскільки в даному дослідженні стоїть задача пошуку закономірностей для багатовимірного випадку, то доцільно використовувати дерева рішень.

У свою чергу, дерева рішень будуються за такими методами:

– Бустинг;

– Багінг;

– Стекінг [12].

Найбільш ефективним з них, як відомо, є бустинг.

Бустинг – це ансамблевий мета-алгоритм машинного навчання передусім для зменшення зсуву а також і дисперсії у навчанні з учителем, та сімейство алгоритмів машинного навчання, які перетворюють слабких учнів на сильних [13].

Багінг – це мета-алгогритм композиційного навчання, призначений для поліпшення стабільності і точності алгоритмів машинного навчання, що використовуються в статистичній класифікації та регресії. Алгоритм також зменшує дисперсію і допомагає уникати перенавчання. Хоча він зазвичай застосовується до методів навчання машин на основі дерев рішень, його можна використовувати з будь-яким видом методу [14].

Стекінг (Stacked Generalization або Stacking) – один з найпопулярніших способів ансамблювання алгоритмів, тобто. використання декількох алгоритмів для вирішення однієї з задач машинного навчання. Він використовується для об'єднання інформації з декількох прогностичних моделей для створення нової моделі. Часто укладені моделі (модель другого рівня) перевершують кожну з окремих моделей завдяки своєму згладжуючому характеру і здатності виділяти кожну базову модель, де вона найкраще виконує роботу, і дискредитувати кожну базову модель, де вона працює погано. З цієї причини стекінг є найбільш ефективним, коли базові моделі значно відрізняються [15].

У свою чергу, для автоматизації бустингу існують три найбільш потужні бібліотеки:

– xgboost;

– catboost;
– lightgbm.

Дані бібліотеки допомагають вирішити багато задач, в тому числі і аналізу [16].

Враховуючи великий розмір дата сету, велику кількість ознак та обмежені обчислювальні можливості, пропонується застосовувати бібліотеку lightgbm. Основними перевагами цієї бібліотеки є:

– Швидке навчання і більш висока ефективність;
– Зменшення використання пам’яті;
– Краща точність прогнозування;
– Можливість обробки великомасштабних даних [15].

1.3 Вибір оптимальних інформаційних технологій

Python, як і буда-яка інша мова програмування, має свої відмінні риси. Отже, можна виділити наступні:

– Простота використання та універсальність.

Python легкий у вивченні та швидкий для роботи у ньому. Це вимагає менших зусиль для написання програми з використанням Python, ніж інші мови, такі як C ++ або Java. Ви можете зробити більше з меншим кодом за коротший проміжок часу. Python економить багато часу, оскільки дозволяє людям проводити час, роблячи фактичні дослідження, замість того, щоб постійно читати документацію, перш ніж досягати фактичної мети. Іншими словами, ви можете орієнтуватися на свою головну мету, а не на інструмент, який ви використовуєте.

– Надзвичайно стабільні бібліотеки з великою підтримкою.

Багатогранність Python також означає, що існує велика кількість бібліотек. Існує понад 125 000 сторонніх бібліотек Python, які роблять Python більш корисним для конкретних цілей, включаючи дослідження. Бібліотеки на зразок NumPy, SciPy, Pandas і matplotlib існують вже давно, надзвичайно

доглянуті, оптимізовані, готові до розробок та добре задокументовані. Спільнота програмістів Python - одна з найкращих у світі; він також великий і дуже активний. У разі будь-якого питання або проблеми, є багато людей, які можуть допомогти.

Одним з головних факторів, завдяки яким Python є таким популярним, є його універсальність. Ви можете використовувати його не тільки для досліджень, але і для веб-розробки, обробки тексту, AI, машинного навчання тощо.

– Python - це динамічна мова.

Python – інтерпретована, об'єктно-орієнтована мова програмування високого рівня з динамічною семантикою. Він має вбудовані структури даних високого рівня, поєднані з динамічним набором тексту та динамічним прив’язкою. Багато програмістів закохуються в Python, оскільки це сприяє підвищенню продуктивності. Під час використання Python немає необхідності у компілюванні. Це означає, що можна бути продуктивними відразу, що допомагає при первинному аналізі даних. Як результат, підхід Python до розробки програмного забезпечення є більш ітеративним.

– Зв'язаний з великою кількістю існуючої інфраструктури.

Спільні документи, що містять живий код, рівняння, візуалізації та текст розповіді. Цей опис звучить дуже знайомо кожному досліднику. Це також знайоме кожному розробнику Python, який використовує ноутбук Jupyter – веб-додаток з відкритим кодом.

Python спільно з The Jupyter Notebook надзвичайно корисний для наукових даних, машинного навчання та досліджень. Це дозволяє людям писати надзвичайно виразний код. Це не тільки допомагає представити результати та вбудовувати результати та графіки в той самий документ, що і ваш код, але також є величезною допомогою, коли справа стосується проведення семінарів, повідомлень про ваші результати та підготовки людей до роботи з Python.

Навколо ноутбука Jupyter пропонується безліч послуг, наприклад, Google Colaboratory, яка пропонує безкоштовні хмарні обчислення, включаючи доступ до високопродуктивних графічних процесорів для запуску ваших ноутбуків Jupyter. Більше того, Google Colab зв'язується безпосередньо з додатками Google Drive, що дозволяє зберігати всі набори даних та ноутбуки у хмарі Google Drive.

– C та C ++ реалізації.

Шукаючи плюси і мінуси Python, ви можете виявити, що Python - це повільна мова програмування. Це може бути повільніше, ніж інші мови, але для цього є деякі рішення. Розробники можуть реалізовувати алгоритми в C та C ++ і піддавати функції Python за допомогою Cython. Cython поєднує простоту використання Python та швидкість нативного коду. По суті, це Python зі швидкістю C. Cython допомагає розробникам отримати перевагу надзвичайно ефективних алгоритмів (низькорівнева реалізація в C / C ++), не відмовляючись від існуючої екосистеми Python та сприятливого для початківців синтаксису.

Для ефективного використання мови Python користуватимемося дистрибутивом Anaconda. Anaconda – це пакетний менеджер Python-дистрибутив, який містить в собі більше 400 основних бібліотек з відкритим вихідним кодом для наукових і інженерних розрахунків. Із встановлених бібліотек можна відзначити numpy, scikit-learn, scipy,pandas, а також Jupyter Notebooks.

Також у Anaconda є декілька інших переваг :

– Якщо потрібні додаткові пакети після установки, можна використовувати менеджер пакетів conda або pip для установки цих пакетів. Це дуже зручно, так як не потрібно буде турбуватися через сумісності одних пакетів з іншими. Також Conda спрощує перехід між Python 2 і 3.

– Anaconda поставляється разом з Spyder - інтерактивної IDE для наукових розрахунків на мові Python. Дана IDE дозволяє писати, редагувати і тестувати код. Spyder пропонує перегляд і редагування змінних за допомогою

GUI, динамічну інтроспекцію коду, знаходження помилок на льоту і багато іншого. Також, за необхідності, можна інтегрувати Anaconda з іншими Python IDE, включаючи PyCharm і Atom.

1.4 Висновки

У цьому розділі проаналізовано інформаційні технології аналізу рівня знань іноземної мови студентів закладів вищої освіти. Дослідження показали, що існуючі методи недостатньо приділяють увагу виявленню закономірностей щодо факторів, які впливають на підвищення рівня знань іноземної мови. Тому в наступних розділах пропонується розробити такий метод.

2 РОЗРОБКА МЕТОДУ ДОСЛІДЖЕННЯ ДАНИХ

2.1 Підготовка даних

Для повного розуміння даних проведемо розвідувальний аналіз. Розглянемо кожен семестр окремо. Але спочатку імпортуємо усі потрібні пакети, та підготуємо дані. На рисунку 2.1 зображено основні пакети, що буде використано у програмі.

```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import seaborn as sns
#import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
sns.set_style('darkgrid')
```

Рисунок 2.1 – Завантаження пакетів даних

Завантажимо перший датасет з даними за 2 семестр навчання. Для цього використаємо бібліотеку pandas та її функцію read_csv() (рис 2.2) [17].

```
df=pd.read_csv('2term.csv')
```

Рисунок 2.2 – Завантаження файлу з даними за 2 семестр

Для того щоб зрозуміти подальші кроки обробки, переглянемо, що саме знаходиться у файлі. На рисунках 2.3, 2.4 продемонстровано приклад даних у датасеті за допомогою функції head().

```
df.head(10)
```

	№ групи	Середній бал	Іноземна мова	ECTS	Трад.	Латинська мова та медична термінологія	ECTS.1	Трад..1	Медична біологія	ECTS.2	...	Трад..3	Основи економічних теорій	ECTS.4	Трад..4	Основи психології та педагогіки	E
0	1-а	3,25	135.0	E	3.0	155.0	D	3.0	149.0	D	...	4.0	182.0	A	зараховано	155.0	
1	10-б	3,75	156.0	D	3.0	170.0	B	4.0	162.0	C	...	4.0	189.0	A	зараховано	166.0	
2	17-а	3,25	124.0	E	3.0	131.0	E	3.0	140.0	E	...	4.0	182.0	A	зараховано	137.0	
3	23-б	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
4	15-а	3,75	157.0	D	3.0	171.0	B	4.0	162.0	C	...	4.0	182.0	A	зараховано	120.0	
5	9-б	4	168.0	C	4.0	193.0	A	5.0	130.0	E	...	4.0	178.0	B	зараховано	145.0	
6	6-б	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
7	7-а	3	143.0	D	3.0	136.0	E	3.0	155.0	D	...	3.0	178.0	B	зараховано	160.0	
8	7-а	4	168.0	C	4.0	183.0	A	5.0	155.0	D	...	4.0	193.0	A	зараховано	160.0	
9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	129.0	E	зараховано	NaN	

10 rows × 23 columns

Рисунок 2.3 – Приклад даних за 2 семестр

```
df.head(10)
```

тинська мова та едична нологія	ECTS.1	Трад..1	Медична біологія	ECTS.2	...	Трад..3	Основи економічних теорій	ECTS.4	Трад..4	Основи психології та педагогіки	ECTS.5	Трад..5	Українська мова (за професійним спрямуванням)	ECTS.6	Трад..6
155.0	D	3.0	149.0	D	...	4.0	182.0	A	зараховано	155.0	D	зараховано	NaN	NaN	NaN
170.0	B	4.0	162.0	C	...	4.0	189.0	A	зараховано	166.0	C	зараховано	NaN	NaN	NaN
131.0	E	3.0	140.0	E	...	4.0	182.0	A	зараховано	137.0	E	зараховано	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
171.0	B	4.0	162.0	C	...	4.0	182.0	A	зараховано	120.0	E	зараховано	NaN	NaN	NaN
193.0	A	5.0	130.0	E	...	4.0	178.0	B	зараховано	145.0	D	зараховано	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
136.0	E	3.0	155.0	D	...	3.0	178.0	B	зараховано	160.0	C	зараховано	NaN	NaN	NaN
183.0	A	5.0	155.0	D	...	4.0	193.0	A	зараховано	160.0	C	зараховано	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	...	NaN	129.0	E	зараховано	NaN	NaN	NaN	NaN	NaN	NaN

Рисунок 2.4 – Приклад даних за 2 семестр (продовження)

З рисунку можна побачити:

– Доступно 23 колонки з даними, але кожен предмет має 3 колонки, що несуть у собі схожу інформацію, а саме: оцінка за предмет за 200 бальною

шкалою, за системою ECTS та традиційною 5-ти бальною шкалою для іспитів та значення «зараховано», «не зараховано» для заліків;

– Перша колонка містить номер групи;
– Друга колонка містить середній бал студента;
– Деякі значення відсутні.

На рисунку 2.5 продемонстровано використання функції info(), що надає можливість розглянути типи даних та кількість не нульових значень датасету.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 426 entries, 0 to 425
Data columns (total 23 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   № групи                                         416 non-null    object
 1   Середній бал                                    378 non-null    object
 2   Іноземна мова                                   378 non-null    float64
 3   ECTS                                            378 non-null    object
 4   Трад.                                           378 non-null    float64
 5   Латинська мова та медична термінологія          378 non-null    float64
 6   ECTS.1                                          378 non-null    object
 7   Трад..1                                         378 non-null    float64
 8   Медична біологія                                378 non-null    float64
 9   ECTS.2                                          378 non-null    object
 10  Трад..2                                         378 non-null    float64
 11  Медична та біологічна фізика                    378 non-null    float64
 12  ECTS.3                                          378 non-null    object
 13  Трад..3                                         378 non-null    float64
 14  Основи економічних теорій                       383 non-null    float64
 15  ECTS.4                                          383 non-null    object
 16  Трад..4                                         383 non-null    object
 17  Основи психології та педагогіки                 380 non-null    float64
 18  ECTS.5                                          380 non-null    object
 19  Трад..5                                         380 non-null    object
 20  Українська мова (за професійним спрямуванням)   2 non-null      float64
 21  ECTS.6                                          2 non-null      object
 22  Трад..6                                         2 non-null      float64
dtypes: float64(12), object(11)
memory usage: 76.7+ KB
```

Рисунок 2.5 – Типи даних та кількість не нульових значень

Вже на цьому етапі можна зробити висновок, що даними з предмету «Українська мова (за професійним спрямування)» можна знехтувати. Також, важливо відзначити, що даних не так і багато, це може вплинути на точність аналізу.

Виведемо основні статистичні параметри даних датасету за допомогою функції describe() (рис. 2.6).

```
df.describe()
```

	Іноземна мова	Трад.	Латинська мова та медична термінологія	Трад..1	Медична біологія	Трад..2	Медична та біологічна фізика	Трад..3	Основи економічних теорій	Основи психології та педагогіки	Українська мова (за професійним спрямуванням)	Трад..6
count	378.000000	378.000000	378.000000	378.000000	378.000000	378.000000	378.000000	378.000000	383.000000	380.000000	2.000000	2.0
mean	157.431217	3.783069	167.642857	4.068783	158.256614	3.767196	166.224868	3.989418	179.138381	158.828947	188.000000	5.0
std	23.838799	0.833979	21.819660	0.852849	17.701182	0.720141	17.252831	0.767299	11.116431	19.146834	8.485281	0.0
min	62.000000	2.000000	122.000000	3.000000	122.000000	3.000000	122.000000	3.000000	128.000000	120.000000	182.000000	5.0
25%	134.250000	3.000000	150.250000	3.000000	145.000000	3.000000	154.250000	3.000000	175.000000	143.000000	185.000000	5.0
50%	160.000000	4.000000	170.000000	4.000000	160.000000	4.000000	166.500000	4.000000	182.000000	155.000000	188.000000	5.0
75%	180.000000	5.000000	186.000000	5.000000	168.750000	4.000000	180.000000	5.000000	185.000000	172.000000	191.000000	5.0
max	200.000000	5.000000	200.000000	5.000000	200.000000	5.000000	200.000000	5.000000	200.000000	200.000000	194.000000	5.0

Рисунок 2.6 – Основні статистичні параметри

З рисунку видно, що дані, в основному, є збалансованими. Виведемо кількість нульових значень по колонках (рис 2.7).

```
df.isnull().sum()
```

```
№ групи                                          10
Середній бал                                     48
Іноземна мова                                    48
ECTS                                             48
Трад.                                            48
Латинська мова та медична термінологія           48
ECTS.1                                           48
Трад..1                                          48
Медична біологія                                 48
ECTS.2                                           48
Трад..2                                          48
Медична та біологічна фізика                     48
ECTS.3                                           48
Трад..3                                          48
Основи економічних теорій                        43
ECTS.4                                           43
Трад..4                                          43
Основи психології та педагогіки                  46
ECTS.5                                           46
Трад..5                                          46
Українська мова (за професійним спрямуванням)   424
ECTS.6                                          424
Трад..6                                         424
dtype: int64
```

Рисунок 2.7 – Кількість нульових значень по колонках

У всіх колонках, їх майже однакова кількість, окрім останньої. Не дивлячись на те, що даних і так не велика кількість, видалимо стрічки з нульовими значеннями. В подальшому аналізі та моделюванні, вони можуть негативно вплинути на результат.

Ще раз розглянемо заголовки колонок (рис. 2.8).

```
df.columns
```

```
Index(['№ групи', 'Середній бал', 'Іноземна мова', 'ECTS', 'Трад.',
       'Латинська мова та медична термінологія', 'ECTS.1', 'Трад..1',
       'Медична біологія', 'ECTS.2', 'Трад..2', 'Медична та біологічна фізика',
       'ECTS.3', 'Трад..3', 'Основи економічних теорій', 'ECTS.4', 'Трад..4',
       'Основи психології та педагогіки', 'ECTS.5', 'Трад..5',
       'Українська мова (за професійним спрямуванням)', 'ECTS.6', 'Трад..6'],
      dtype='object')
```

Рисунок 2.8 – Заголовки колонок

Як вже було описано вище, кожен предмет має 3 колонки з даними, що несуть однакову інформацію. Тому, зробимо вибірку з даних, з якою буде доцільніше проводити аналіз та моделювання (рис. 2.9).

```
small = df[['№ групи','Латинська мова та медична термінологія','Медична біологія',
            'Медична та біологічна фізика','Основи економічних теорій','Основи психології та педагогіки',
            'Трад.']]
```

```
small.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 378 entries, 0 to 425
Data columns (total 7 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   № групи                               378 non-null    object
 1   Латинська мова та медична термінологія  378 non-null    float64
 2   Медична біологія                      378 non-null    float64
 3   Медична та біологічна фізика          378 non-null    float64
 4   Основи економічних теорій             378 non-null    float64
 5   Основи психології та педагогіки       378 non-null    float64
 6   Трад.                                 378 non-null    float64
dtypes: float64(6), object(1)
memory usage: 23.6+ KB
```

Рисунок 2.9 – Вибірка даних за 2 семестр

Обсяг даних, очевидно, зменшився, приблизно, у 3 рази. Це пришвидшить роботу програми, в цілому. Оцінку з англійської мови обираємо у традиційному форматі, це полегшить прогнозування, так як кількість унікальних значень у цій колонці мала.

На цьому підготовку можна завершити, перейдемо до файлу з даними за 4 семестр та проведемо аналогічні операції з ним.

Завантажимо другий файл з даними за 4 семестр навчання (рис. 2.10).

```
df=pd.read_csv('4term.csv')
```

Рисунок 2.10 – Завантаження файлу з даними за 4 семестр

Переглянемо дані, на рисунках 2.11, 2.12 продемонстровано приклад даних у датасеті.

```
df.head(10)
```

	№ групи	Середній бал	Безпека життєдіяльності та охорона праці	ECTS	Трад.	Біологічна та біоорганічна хімія	ECTS.1	Трад..1	Догляд за хворими	ECTS.2	...	Трад..7	Філософія	ECTS.8	Трад..8	Цив (мед забезп насе.
0	1-а	3	164.0	C	зараховано	146.0	D	3.0	152.0	D	...	3.0	142.0	D	3.0	
1	10-б	4,5	160.0	C	зараховано	151.0	D	3.0	180.0	A	...	5.0	180.0	A	5.0	
2	17-а	3,42	128.0	E	зараховано	138.0	E	3.0	147.0	D	...	3.0	148.0	D	3.0	
3	23-б	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	
4	15-а	4,42	180.0	A	зараховано	163.0	C	4.0	192.0	A	...	4.0	164.0	C	4.0	
5	9-б	3,83	140.0	E	зараховано	133.0	E	3.0	189.0	A	...	3.0	134.0	E	3.0	
6	6-б	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	
7	7-а	3,33	160.0	C	зараховано	127.0	E	3.0	180.0	A	...	3.0	143.0	D	3.0	
8	7-а	4	176.0	B	зараховано	138.0	E	3.0	185.0	A	...	4.0	168.0	C	4.0	
9	14-б	3,42	168.0	C	зараховано	128.0	E	3.0	160.0	C	...	3.0	168.0	C	4.0	

10 rows × 35 columns

Рисунок 2.11 – Приклад даних за 4 семестр

```
df.head(10)
```

Біологічна та біоорганічна хімія	ECTS.1	Трад..1	Догляд за хворими	ECTS.2	...	Трад..7	Філософія	ECTS.8	Трад..8	Цивільний захист. Основи медичного забезпечення населення і військ.	ECTS.9	Трад..9	Догляд за хворими (практика з педіатрії)	ECTS.10	Трад..10
146.0	D	3.0	152.0	D	...	3.0	142.0	D	3.0	175.0	B	зараховано	NaN	NaN	NaN
151.0	D	3.0	180.0	A	...	5.0	180.0	A	5.0	185.0	A	зараховано	NaN	NaN	NaN
138.0	E	3.0	147.0	D	...	3.0	148.0	D	3.0	125.0	E	зараховано	166.0	C	4.0
NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
163.0	C	4.0	192.0	A	...	4.0	164.0	C	4.0	160.0	C	зараховано	171.0	B	4.0
133.0	E	3.0	189.0	A	...	3.0	134.0	E	3.0	135.0	E	зараховано	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
127.0	E	3.0	180.0	A	...	3.0	143.0	D	3.0	185.0	A	зараховано	NaN	NaN	NaN
138.0	E	3.0	185.0	A	...	4.0	168.0	C	4.0	170.0	B	зараховано	NaN	NaN	NaN
128.0	E	3.0	160.0	C	...	3.0	168.0	C	4.0	185.0	A	зараховано	172.0	B	4.0

Рисунок 2.12 – Приклад даних за 4 семестр (продовження)

Відмінність від першого файлу у обсязі даних – доступно 35 колонок, за рахунок більшої кількості предметів. Але структура файлу та ж сама.

Розглянемо типи даних та кількість не нульових значень датасету 4 семестру (рис. 2.13, 2.14).

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 426 entries, 0 to 425
Data columns (total 35 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   № групи                                                 422 non-null    object
 1   Середній бал                                            383 non-null    object
 2   Безпека життєдіяльності та охорона праці                382 non-null    float64
 3   ECTS                                                    382 non-null    object
 4   Трад.                                                   382 non-null    object
 5   Біологічна та біоорганічна хімія                        382 non-null    float64
 6   ECTS.1                                                  382 non-null    object
 7   Трад..1                                                 382 non-null    float64
 8   Догляд за хворими                                       382 non-null    float64
 9   ECTS.2                                                  382 non-null    object
 10  Трад..2                                                 382 non-null    float64
 11  Іноземна мова за професійним спрямуванням               382 non-null    float64
 12  ECTS.3                                                  382 non-null    object
 13  Трад..3                                                 382 non-null    float64
 14  Логіка, формальна логіка                                380 non-null    float64
 15  ECTS.4                                                  380 non-null    object
 16  Трад..4                                                 380 non-null    object
```

Рисунок 2.13 – Типи даних та кількість не нульових значень

```
 17  Медична інформатика                                                  383 non-null    float64
 18  ECTS.5                                                               383 non-null    object
 19  Трад..5                                                              383 non-null    float64
 20  Фізичне виховання                                                    382 non-null    float64
 21  ECTS.6                                                               382 non-null    object
 22  Трад..6                                                              382 non-null    object
 23  Фізіологія                                                           382 non-null    float64
 24  ECTS.7                                                               382 non-null    object
 25  Трад..7                                                              382 non-null    float64
 26  Філософія                                                            382 non-null    float64
 27  ECTS.8                                                               382 non-null    object
 28  Трад..8                                                              382 non-null    float64
 29  Цивільний захист. Основи медичного забезпечення населення і військ.  389 non-null    float64
 30  ECTS.9                                                               389 non-null    object
 31  Трад..9                                                              389 non-null    object
 32  Догляд за хворими (практика з педіатрії)                             126 non-null    float64
 33  ECTS.10                                                              126 non-null    object
 34  Трад..10                                                             126 non-null    float64
dtypes: float64(18), object(17)
memory usage: 116.6+ KB
```

Рисунок 2.14 – Типи даних та кількість не нульових значень (продовження)

Кількість рядків така ж, як у першому файлі, це добре, тому що оцінки у всіх файлах мають відповідати тим самим студентам.

Виведемо основні статистичні параметри даних датасету за допомогою функції describe() (рис. 2.15).

```
df.describe()
```

	Безпека життєдіяльності та охорона праці	Біологічна та біоорганічна хімія	Трад..1	Догляд за хворими	Трад..2	Іноземна мова за професійним спрямуванням	Трад..3	Логіка, формальна логіка	Медична інформатика	Трад..5	Фізичне виховання	Ф
count	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000	380.000000	383.000000	383.000000	382.000000	38
mean	172.680628	151.814136	3.531414	177.476440	4.473822	158.884817	3.835079	157.452632	169.501305	4.125326	173.963351	16
std	14.484586	19.339609	0.715635	12.734446	0.613150	25.631598	0.855137	18.472553	16.420501	0.779287	9.586940	1
min	128.000000	122.000000	3.000000	130.000000	3.000000	122.000000	3.000000	120.000000	124.000000	3.000000	141.000000	12
25%	164.000000	135.000000	3.000000	169.000000	4.000000	132.000000	3.000000	142.000000	160.000000	4.000000	167.000000	15
50%	172.000000	148.000000	3.000000	180.000000	5.000000	160.000000	4.000000	160.000000	170.000000	4.000000	174.000000	16
75%	184.000000	164.000000	4.000000	187.000000	5.000000	182.000000	5.000000	169.000000	182.000000	5.000000	180.000000	18
max	200.000000	200.000000	5.000000	200.000000	5.000000	200.000000	5.000000	191.000000	200.000000	5.000000	200.000000	20

Рисунок 2.15 – Основні статистичні параметри

Дані, в основному, є збалансованими. Виведемо кількість нульових значень по колонках (рис 2.16).

```
df.isnull().sum()
```

```
№ групи                                                          4
Середній бал                                                    43
Безпека життєдіяльності та охорона праці                        44
ECTS                                                            44
Трад.                                                           44
Біологічна та біоорганічна хімія                                44
ECTS.1                                                          44
Трад..1                                                         44
Догляд за хворими                                               44
ECTS.2                                                          44
Трад..2                                                         44
Іноземна мова за професійним спрямуванням                       44
ECTS.3                                                          44
Трад..3                                                         44
Логіка, формальна логіка                                        46
ECTS.4                                                          46
Трад..4                                                         46
Медична інформатика                                             43
ECTS.5                                                          43
Трад..5                                                         43
Фізичне виховання                                               44
ECTS.6                                                          44
Трад..6                                                         44
Фізіологія                                                      44
ECTS.7                                                          44
Трад..7                                                         44
Філософія                                                       44
ECTS.8                                                          44
Трад..8                                                         44
Цивільний захист. Основи медичного забезпечення населення і військ.    37
ECTS.9                                                          37
Трад..9                                                         37
Догляд за хворими (практика з педіатрії)                       300
ECTS.10                                                        300
Трад..10                                                       300
dtype: int64
```

Рисунок 2.16 – Кількість нульових значень по колонках

Так само, видалимо стрічки з пропущеними значеннями, та знехтуємо останнім предметом – велика кількість пропущених значень.

Тому вибірка з 2 файлу матиме вигляд продемонстрований на рисунку 2.17.

```
small = df[['№ групи','Безпека життєдіяльності та охорона праці','Біологічна та біоорганічна хімія',
            'Догляд за хворими','Логіка, формальна логіка','Медична інформатика',
            'Фізичне виховання','Фізіологія',
            'Цивільний захист. Основи медичного забезпечення населення і військ.','Трад..3']]
```

```
small.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 380 entries, 0 to 425
Data columns (total 10 columns):
 #   Column                                                        Non-Null Count  Dtype
---  ------                                                        --------------  -----
 0   № групи                                                       380 non-null    object
 1   Безпека життєдіяльності та охорона праці                      380 non-null    float64
 2   Біологічна та біоорганічна хімія                              380 non-null    float64
 3   Догляд за хворими                                             380 non-null    float64
 4   Логіка, формальна логіка                                      380 non-null    float64
 5   Медична інформатика                                           380 non-null    float64
 6   Фізичне виховання                                             380 non-null    float64
 7   Фізіологія                                                    380 non-null    float64
 8   Цивільний захист. Основи медичного забезпечення населення і військ.  380 non-null    float64
 9   Трад..3                                                       380 non-null    float64
dtypes: float64(9), object(1)
memory usage: 32.7+ KB
```

Рисунок 2.17 – Вибірка даних за 4 семестр

Наступним завантажимо файл з даними за 6 семестр навчання (рис 2.18). Проведемо підготовку з ним.

```
df=pd.read_csv('6term.csv')
```

Рисунок 2.18 – Завантаження файлу з даними за 6 семестр

На рисунку 2.19 продемонстровано приклад даних у датасеті.

```
df.head(10)
```

	№ групи	Середній бал	Військова гігієна	ECTS	Трад.	Загальна хірургія	ECTS.1	Трад..1	Іноземна мова	ECTS.2	...	Трад..6	Радіологія	ECTS.7	Трад..7	Сестринська практика	ECT
0	1-а	3,4	166.0	C	4.0	174.0	B	4.0	144.0	D	...	3.0	137.0	E	3.0	188.0	
1	10-б	4,2	188.0	A	5.0	169.0	C	4.0	186.0	A	...	4.0	160.0	C	4.0	180.0	
2	17-а	3,1	122.0	E	3.0	149.0	D	3.0	133.0	E	...	3.0	142.0	D	3.0	160.0	
3	12-б	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	0.0	
4	15-а	4,5	185.0	A	5.0	166.0	C	4.0	174.0	B	...	4.0	180.0	A	5.0	191.0	
5	9-б	3	150.0	D	3.0	154.0	D	3.0	148.0	D	...	3.0	133.0	E	3.0	151.0	
6	12-б	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	0.0	
7	7-а	3,3	137.0	E	3.0	161.0	C	4.0	152.0	D	...	3.0	126.0	E	3.0	180.0	
8	7-а	3,5	149.0	D	3.0	160.0	C	4.0	187.0	A	...	3.0	148.0	D	3.0	184.0	
9	14-б	3,3	134.0	E	3.0	184.0	A	5.0	128.0	E	...	3.0	132.0	E	3.0	160.0	

10 rows × 32 columns

Рисунок 2.19 – Приклад даних за 6 семестр

Доступно 32 колонки і зновупропущені значення. Результат використання функції info() зображено на рисунку 2.20.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 426 entries, 0 to 425
Data columns (total 32 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   № групи                               426 non-null    object
 1   Середній бал                          423 non-null    object
 2   Військова гігієна                     412 non-null    float64
 3   ECTS                                  412 non-null    object
 4   Трад.                                 412 non-null    float64
 5   Загальна хірургія                     410 non-null    float64
 6   ECTS.1                                410 non-null    object
 7   Трад..1                               410 non-null    float64
 8   Іноземна мова                         410 non-null    float64
 9   ECTS.2                                410 non-null    object
 10  Трад..2                               410 non-null    float64
 11  Патоморфологія                        410 non-null    float64
 12  ECTS.3                                410 non-null    object
 13  Трад..3                               410 non-null    float64
 14  Патофізіологія                        410 non-null    float64
 15  ECTS.4                                410 non-null    object
 16  Трад..4                               410 non-null    float64
 17  Пропедевтика внутрішньої  медицини    410 non-null    float64
 18  ECTS.5                                410 non-null    object
 19  Трад..5                               410 non-null    float64
 20  Пропедевтика педіатрії                410 non-null    float64
 21  ECTS.6                                410 non-null    object
 22  Трад..6                               410 non-null    float64
 23  Радіологія                            410 non-null    float64
 24  ECTS.7                                410 non-null    object
 25  Трад..7                               410 non-null    float64
 26  Сестринська практика                  422 non-null    float64
 27  ECTS.8                                422 non-null    object
 28  Трад..8                               422 non-null    float64
 29  Фармакологія                          410 non-null    float64
 30  ECTS.9                                410 non-null    object
 31  Трад..9                               410 non-null    object
dtypes: float64(19), object(13)
memory usage: 106.6+ KB
```

Рисунок 2.20 – Типи даних та кількість не нульових значень

Кількість рядків у трьох файлах однакова, також, у цьому файлі найменше пропущених значень.

Виведемо основні статистичні параметри даних датасету (рис. 2.21)

```
df.describe()
```

	Військова гігієна	Трад.	Загальна хірургія	Трад..1	Іноземна мова	Трад..2	Патоморфологія	Трад..3	Патофізіологія	Трад..4	Пропедевтика внутрішньої медицини	
count	412.000000	412.000000	410.000000	410.000000	410.000000	410.000000	410.000000	410.000000	410.00000	410.000000	410.000000	410
mean	163.157767	3.917476	169.463415	4.139024	160.363415	3.819512	154.780488	3.600000	149.40000	3.536585	157.080488	3
std	18.011942	0.720223	16.655575	0.650513	25.463760	0.871505	20.987611	0.779317	27.85381	0.806379	20.522177	0
min	122.000000	3.000000	0.000000	2.000000	36.000000	2.000000	99.000000	2.000000	0.00000	2.000000	42.000000	2
25%	154.000000	3.000000	163.000000	4.000000	141.000000	3.000000	137.000000	3.000000	131.00000	3.000000	144.000000	3
50%	163.500000	4.000000	170.000000	4.000000	162.000000	4.000000	151.000000	3.000000	146.00000	3.000000	160.000000	4
75%	177.000000	4.000000	180.000000	5.000000	182.000000	5.000000	168.000000	4.000000	168.00000	4.000000	169.000000	4
max	200.000000	5.000000	200.000000	5.000000	200.000000	5.000000	200.000000	5.000000	200.00000	5.000000	200.000000	5

Рисунок 2.21 – Основні статистичні параметри

Виведемо кількість нульових значень скориставшись функціями isnull() та sum() (рис. 2.22, 2.23).

```
df.isnull().sum()
```

```
Прізвище, ім'я та по батькові          0
№ групи                                0
Середній бал                           3
Військова гігієна                     14
ECTS                                  14
Трад.                                 14
Загальна хірургія                     16
ECTS.1                                16
Трад..1                               16
Іноземна мова                         16
ECTS.2                                16
Трад..2                               16
Патоморфологія                        16
ECTS.3                                16
Трад..3                               16
Патофізіологія                        16
ECTS.4                                16
Трад..4                               16
Пропедевтика внутрішньої  медицини    16
ECTS.5                                16
Трад..5                               16
Пропедевтика педіатрії                16
ECTS.6                                16
Трад..6                               16
```

Рисунок 2.22 – Кількість нульових значень по колонках

```
Радіологія                          16
ECTS.7                              16
Трад..7                             16
Сестринська практика                 4
ECTS.8                               4
Трад..8                              4
Фармакологія                        16
ECTS.9                              16
Трад..9                             16
```

Рисунок 2.23 – Кількість нульових значень по колонках (продовження)

В файлі з оцінками за 6 семестр найменша кількість пропущених значень. Тому і вибірка буде найбільшою (рис. 2.24).

```
small = df[['№ групи','Військова гігієна','Загальна хірургія',
            'Патоморфологія','Патофізіологія','Пропедевтика внутрішньої  медицини',
            'Пропедевтика педіатрії','Сестринська практика','Фармакологія','Трад..2']]
```

```
small.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 426 entries, 0 to 425
Data columns (total 10 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   № групи                             426 non-null    object
 1   Військова гігієна                   412 non-null    float64
 2   Загальна хірургія                   410 non-null    float64
 3   Патоморфологія                      410 non-null    float64
 4   Патофізіологія                      410 non-null    float64
 5   Пропедевтика внутрішньої  медицини  410 non-null    float64
 6   Пропедевтика педіатрії              410 non-null    float64
 7   Сестринська практика                422 non-null    float64
 8   Фармакологія                        410 non-null    float64
 9   Трад..2                             410 non-null    float64
dtypes: float64(9), object(1)
memory usage: 33.4+ KB
```

Рисунок 2.24 – Вибірка даних за 6 семестр

Останнім кроком є об’єднання усіх датасетів та додання до них балів за ЄДКІ з англійської мови. На рисунку 2.25 показано приклад даних оцінювання ЄДКІ.

```
df3.head()
```

	№	Ідентиф.	Група	Бал	Тест
0	1	359020011	11	191	53
1	2	359020117	11	184	46
2	3	359020679	11	215	78
3	4	359020833	11	205	67
4	5	359022790	11	222	85

Рисунок 2.25 – Приклад даних оцінювання ЄДКІ

З цього файлу ми приєднуємо відсоток за тест, до балів з усіх інших дисциплін. Колонки, що входять до повного датасету продемонстровано на рисунку 2.26.

```
cutdatafull=datafull[['№ групи','Латинська мова та медична термінологія',
                      'Медична біологія','Медична та біологічна фізика',
                      'Основи економічних теорій','Основи психології та педагогіки',
                      'Іноземна мова_х','Безпека життєдіяльності та охорона праці',
                      'Біологічна та біоорганічна хімія',  'Догляд за хворими',
                      'Логіка, формальна логіка','Медична інформатика',
                      'Фізичне виховання','Фізіологія',
                      'Цивільний захист. Основи медичного забезпечення населення і військ.
                      'Іноземна мова за професійним спрямуванням', 'Військова гігієна',
                      'Загальна хірургія','Патоморфологія','Патофізіологія',
                      'Пропедевтика внутрішньої  медицини', 'Пропедевтика педіатрії',
                      'Сестринська практика','Фармакологія','Іноземна мова_у', 'Тест']]
```

Рисунок 2.26 – Колонки повного датасету

Остання колонка має багато унікальних значень, так як в ній міститься відсоток правильних відповідей за тест. МОЗ України встановлено, мінімальний критерій успішного складання є 60,5 % (Наказ МОЗ України від 05.10.2016 №1043) [18]. Тому відсотки було перетворено у 5-ти бальну і систему і будемо вважати, що усі результати 3 і вище є успішними. Код для цього перетворення продемонстровано на рисунку 2.27.

```
cutdatafull['Тест']=cutdatafull['Тест'].mask(cutdatafull['Тест'] >= 90,5)
cutdatafull['Тест']=cutdatafull['Тест'].mask(cutdatafull['Тест'] >= 75,4)
cutdatafull['Тест']=cutdatafull['Тест'].mask(cutdatafull['Тест'] >= 61,3)
cutdatafull['Тест']=cutdatafull['Тест'].mask(cutdatafull['Тест'] >= 6,2)
```

Рисунок 2.27 – Перетворення відсотків у 5-ти бальну систему

На рисунках 2.28 та 2.29 продемонстровано приклад даних у повному датасеті.

```
df.head(10)
```

	№ групи	Середній бал	Військова гігієна	ECTS	Трад.	Загальна хірургія	ECTS.1	Трад..1	Іноземна мова	ECTS.2	...	Трад..6	Радіологія	ECTS.7	Трад..7	Сестринська практика	ECT
0	1-а	3,4	166.0	C	4.0	174.0	B	4.0	144.0	D	...	3.0	137.0	E	3.0	188.0	
1	10-б	4,2	188.0	A	5.0	169.0	C	4.0	186.0	A	...	4.0	160.0	C	4.0	180.0	
2	17-а	3,1	122.0	E	3.0	149.0	D	3.0	133.0	E	...	3.0	142.0	D	3.0	160.0	
3	12-б	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	0.0	
4	15-а	4,5	185.0	A	5.0	166.0	C	4.0	174.0	B	...	4.0	180.0	A	5.0	191.0	
5	9-б	3	150.0	D	3.0	154.0	D	3.0	148.0	D	...	3.0	133.0	E	3.0	151.0	
6	12-б	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	0.0	
7	7-а	3,3	137.0	E	3.0	161.0	C	4.0	152.0	D	...	3.0	126.0	E	3.0	180.0	
8	7-а	3,5	149.0	D	3.0	160.0	C	4.0	187.0	A	...	3.0	148.0	D	3.0	184.0	
9	14-б	3,3	134.0	E	3.0	184.0	A	5.0	128.0	E	...	3.0	132.0	E	3.0	160.0	

10 rows × 32 columns

Рисунок 2.28 – Приклад даних за усі семестри

Біологічна та біоорганічна хімія	Догляд за хворими	...	Військова гігієна	Загальна хірургія	Патоморфологія	Патофізіологія	Пропедевтика внутрішньої медицини	Пропедевтика педіатрії	Сестринська практика	Фармакологія	Іноземна мова_у	Тест
146.0	152.0	...	166.0	174.0	135.0	130.0	142.0	135.0	188.0	147.0	144.0	53
151.0	180.0	...	188.0	169.0	173.0	143.0	165.0	164.0	180.0	161.0	186.0	71
138.0	147.0	...	122.0	149.0	132.0	130.0	126.0	136.0	160.0	140.0	133.0	42
163.0	192.0	...	185.0	166.0	173.0	194.0	161.0	165.0	191.0	187.0	174.0	92
133.0	189.0	...	150.0	154.0	139.0	130.0	132.0	145.0	151.0	141.0	148.0	75

Рисунок 2.29 – Приклад даних за усі семестри (продовження)

Доступно 32 колонки і зновупропущені значення. Результат використання функції info() зображено на рисунку 2.30.

```
datafull.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 426 entries, 0 to 425
Data columns (total 28 columns):
 #   Column                                                    Non-Null Count  Dtype
---  ------                                                    --------------  -----
 1   № групи_x                                                 416 non-null    object
 2   Латинська мова та медична термінологія                    378 non-null    float64
 3   Медична біологія                                          378 non-null    float64
 4   Медична та біологічна фізика                              378 non-null    float64
 5   Основи економічних теорій                                 383 non-null    float64
 6   Основи психології та педагогіки                           380 non-null    float64
 7   Іноземна мова_x                                           378 non-null    float64
 8   № групи_y                                                 422 non-null    object
 9   Безпека життєдіяльності та охорона праці                  382 non-null    float64
 10  Біологічна та біоорганічна хімія                          382 non-null    float64
 11  Догляд за хворими                                         382 non-null    float64
 12  Логіка, формальна логіка                                  380 non-null    float64
 13  Медична інформатика                                       383 non-null    float64
 14  Фізичне виховання                                         382 non-null    float64
 15  Фізіологія                                                382 non-null    float64
 16  Цивільний захист. Основи медичного забезпечення населення і військ.  389 non-null    float64
 17  Іноземна мова за професійним спрямуванням                 382 non-null    float64
 18  № групи                                                   426 non-null    object
 19  Військова гігієна                                         412 non-null    float64
 20  Загальна хірургія                                         410 non-null    float64
 21  Патоморфологія                                            410 non-null    float64
 22  Патофізіологія                                            410 non-null    float64
 23  Пропедевтика внутрішньої  медицини                        410 non-null    float64
 24  Пропедевтика педіатрії                                    410 non-null    float64
 25  Сестринська практика                                      422 non-null    float64
 26  Фармакологія                                              410 non-null    float64
 27  Іноземна мова_y                                           410 non-null    float64
dtypes: float64(24), object(4)
memory usage: 96.5+ KB
```

Рисунок 2.30 – Типи даних та кількість не нульових значень

Кількість рядків у трьох файлах однакова, також, у цьому файлі найменше пропущених значень.

Виведемо основні статистичні параметри даних датасету (рис. 2.31, 2.32)

```
df.describe()
```

	№ групи	Латинська мова та медична термінологія	Медична біологія	Медична та біологічна фізика	Основи економічних теорій	Основи психології та педагогіки	Іноземна мова_x	Безпека життєдіяльності та охорона праці	Біологічна та біоорганічна хімія	Догляд за хворими
count	355.000000	355.000000	355.000000	355.000000	354.000000	355.000000	354.000000	353.000000	355.000000	353.000000
mean	18.673239	167.566197	158.346479	166.411268	179.641243	158.661972	157.559322	172.974504	151.797183	177.436261
std	10.339450	21.859412	17.773082	17.164044	9.956483	18.858785	23.602903	14.249479	19.115424	12.358381
min	1.000000	122.000000	122.000000	122.000000	153.000000	120.000000	122.000000	132.000000	123.000000	142.000000
25%	10.000000	150.000000	145.000000	155.000000	175.000000	143.000000	134.250000	164.000000	135.500000	169.000000
50%	19.000000	169.000000	160.000000	166.000000	182.000000	160.000000	160.000000	172.000000	148.000000	180.000000
75%	28.000000	186.000000	168.500000	180.000000	185.000000	172.000000	180.000000	184.000000	164.000000	186.000000
max	36.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000

8 rows × 26 columns

Рисунок 2.31 – Основні статистичні параметри

Військова гігієна	Загальна хірургія	Патоморфологія	Патофізіологія	Пропедевтика внутрішньої медицини	Пропедевтика педіатрії	Сестринська практика	Фармакологія	Іноземна мова_у	Тест
355.000000	354.000000	355.000000	349.000000	353.000000	354.000000	353.000000	354.000000	352.000000	355.000000
164.208451	170.672316	156.022535	152.790831	159.005666	162.259887	179.518414	158.533898	163.147727	68.701408
17.599000	13.764510	20.608115	22.364389	17.789807	17.885700	10.870902	18.516365	22.816019	16.785987
122.000000	131.000000	99.000000	92.000000	122.000000	122.000000	150.000000	125.000000	108.000000	21.000000
154.000000	163.000000	139.000000	133.000000	147.000000	149.250000	172.000000	144.000000	142.750000	57.000000
164.000000	170.000000	152.000000	148.000000	161.000000	163.000000	180.000000	157.000000	163.000000	71.000000
177.000000	180.000000	169.000000	168.000000	170.000000	176.750000	189.000000	171.000000	184.000000	82.000000
200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	100.000000

Рисунок 2.32 – Основні статистичні параметри

Замінемо усі нульові значення (рис. 2.33).

```
cutdatafull = cutdatafull.replace(0, np.nan)
```

Рисунок 2.33 – Основні статистичні параметри

Потім видалимо пусті значення (рис. 2.34).

```
cutdatafull=cutdatafull.dropna()
```

Рисунок 2.34 – Основні статистичні параметри

В файлі з оцінками за 6 семестр найменша кількість пропущених значень. Тому і вибірка буде найбільшою (рис. 2.35).

```
cutdatafull.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 358 entries, 0 to 426
Data columns (total 26 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   № групи                                                 358 non-null    int32
 1   Латинська мова та медична термінологія                  358 non-null    int32
 2   Медична біологія                                        358 non-null    int32
 3   Медична та біологічна фізика                            358 non-null    int32
 4   Основи економічних теорій                               358 non-null    int32
 5   Основи психології та педагогіки                         358 non-null    int32
 6   Іноземна мова_х                                         358 non-null    int32
 7   Безпека життєдіяльності та охорона праці                358 non-null    int32
 8   Біологічна та біоорганічна хімія                        358 non-null    int32
 9   Догляд за хворими                                       358 non-null    int32
 10  Логіка, формальна логіка                                358 non-null    int32
 11  Медична інформатика                                     358 non-null    int32
 12  Фізичне виховання                                       358 non-null    int32
 13  Фізіологія                                              358 non-null    int32
 14  Цивільний захист. Основи медичного забезпечення населення і військ.  358 non-null    int32
 15  Іноземна мова за професійним спрямуванням               358 non-null    int32
 16  Військова гігієна                                       358 non-null    int32
 17  Загальна хірургія                                       358 non-null    int32
 18  Патоморфологія                                          358 non-null    int32
 19  Патофізіологія                                          358 non-null    int32
 20  Пропедевтика внутрішньої  медицини                      358 non-null    int32
 21  Пропедевтика педіатрії                                  358 non-null    int32
 22  Сестринська практика                                    358 non-null    int32
 23  Фармакологія                                            358 non-null    int32
 24  Іноземна мова_у                                         358 non-null    int32
 25  Тест                                                    358 non-null    int32
dtypes: int32(26)
memory usage: 39.2 KB
```

Рисунок 2.35 – Вибірка даних за 6 семестр

Файли з даними за 2, 4, 6 семестри навчання та загальний файл було успішно підготовлено до наступних етапів роботи, а саме до розвідувального аналізу та моделювання. З кожного файлу було сформовано вибірку, за допомогою яких і було проведено ці етапи.

2.2 Розвідувальний аналіз

Розвідувальний аналіз даних або (EDA) - це розуміння наборів даних шляхом узагальнення їх основних характеристик, часто їх візуально будують. Цей крок дуже важливий, особливо коли ми підходимо до моделювання даних, щоб застосувати машинне навчання. Складання графіку в EDA часто

складається з гістограм, графіків залежностей, кореляційних матриць і тд. На дослідження даних часто потрібно багато часу [10].

Результати розвідувального аналізу не використовуються для вироблення управлінських рішень. Їхнє призначення – допомога в розробці найкращої стратегії поглибленого аналізу, висування гіпотез, уточнення особливостей застосування тих чи інших математичних методів та моделей.

Оскільки найпоширенішою мовою програмування для інтелектуального аналізу і передбачення даних в наш час є мова Python, використаємо її для розвідувального аналізу [19].

Дані датасетів доволі однотипні та зрозумілі – бали по різним дисциплінам, тому візуалізація є мінімальною. Спочатку було побудовано графік кореляції для першого датасету. На рисунках 2.36 і 2.37 бачимо відповідно її код в середовищі Python і її візуальне представлення.

```
corrmat = small.corr()
f, ax = plt.subplots(figsize=(12, 9))
sns.heatmap(corrmat, vmax=.8, square=True);
```

Рисунок 2.36 – Код для побудови кореляційної теплової карти

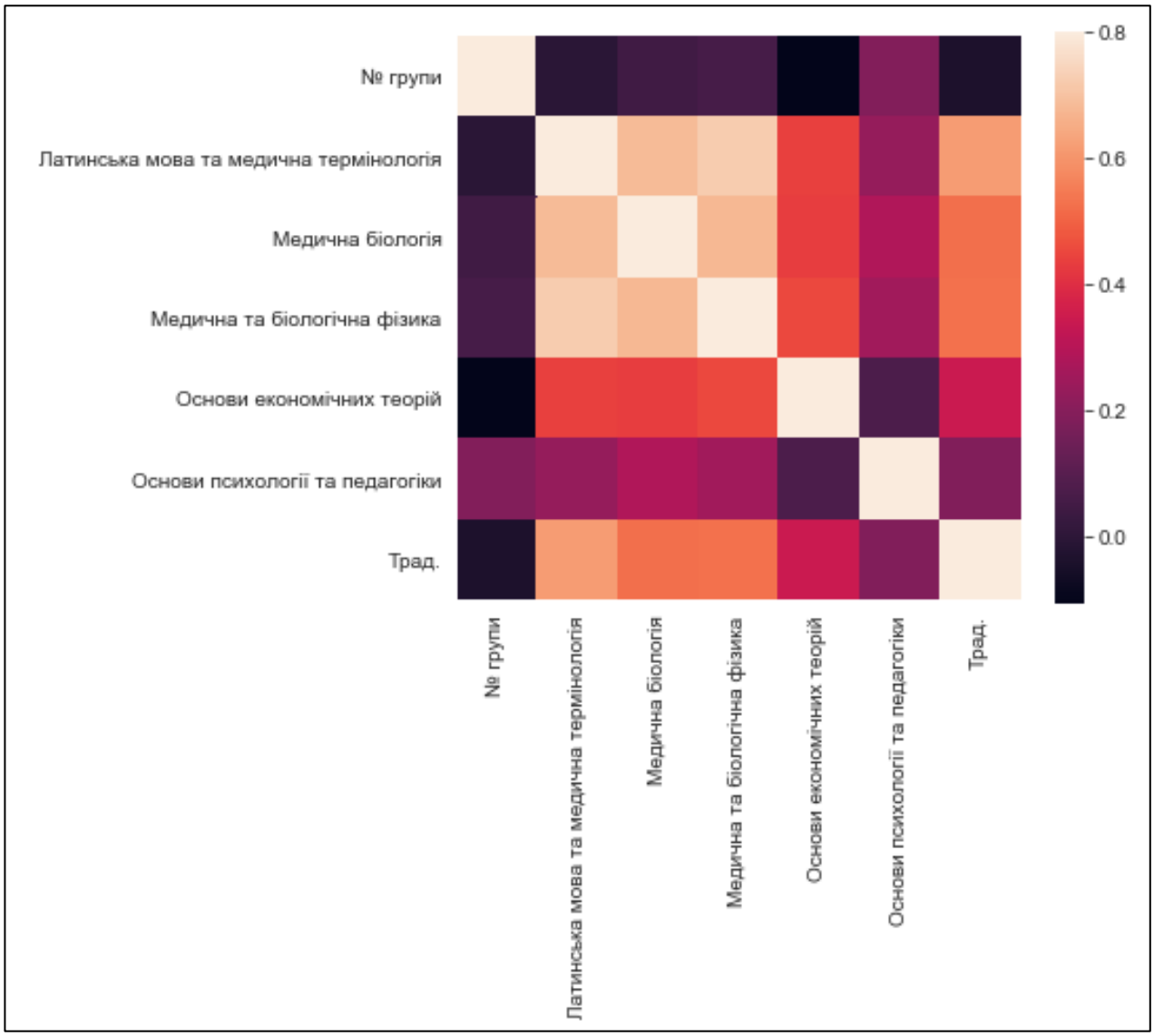

Рисунок 2.37 – Теплова кореляційна карта

На основі цієї карти можна зробити висновок, що існує деяка залежність між першими трьома колонками. Але нас цікавить, що впливає на результати з англійської мови. Найвищу кореляцію з усіх доступних предметів має «Латинська мова та медична термінологія», що є логічним – здатність до вивчення одної мови впливає на вивчення іншим мов.

Далі було побудовано графіки розподілу значень по усіх дисциплінах. Ці графіки та код для їх побудови продемонстровано на рисунках 2.38, 2.39.

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Іноземна мова')
sns.violinplot(y='Іноземна мова',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Латинська мова та медична термінологія')
sns.violinplot(y='Латинська мова та медична термінологія',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Медична біологія ')
sns.violinplot(y='Медична біологія',data=small,palette='spring',linewidth=3)
plt.show()
```

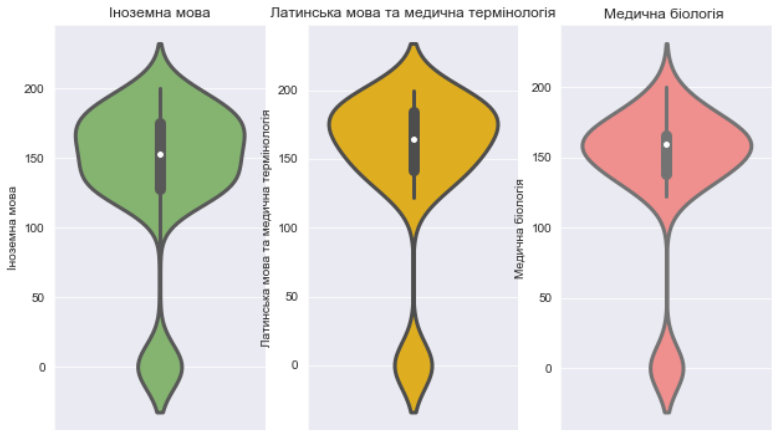

Рисунок 2.38 – Графіки розподілу значень перших трьох дисциплін

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Медична та біологічна фізика')
sns.violinplot(y='Медична та біологічна фізика',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Основи економічних теорій')
sns.violinplot(y='Основи економічних теорій',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Основи психології та педагогіки')
sns.violinplot(y='Основи психології та педагогіки',data=small,palette='spring',linewidth=3)
plt.show()
```

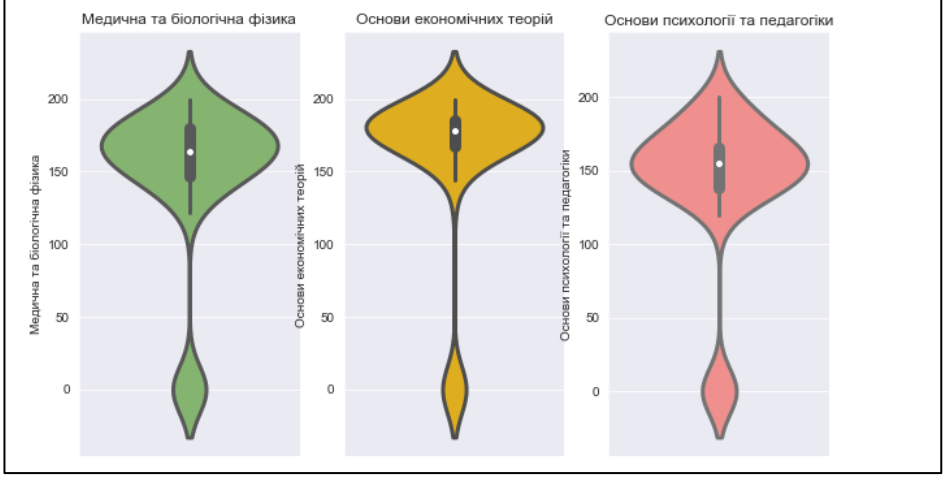

Рисунок 2.39 – Графіки розподілу значень останніх трьох дисциплін

На основі графіків розподілу можна зробити висновок, що найбільш схожим до розподілу значень колонки «Іноземна мова» є значення колонки «Латинська мова та медична термінологія».

Також побудуємо гістограму розподілу дисципліни «Іноземна мова» (рис. 2.40).

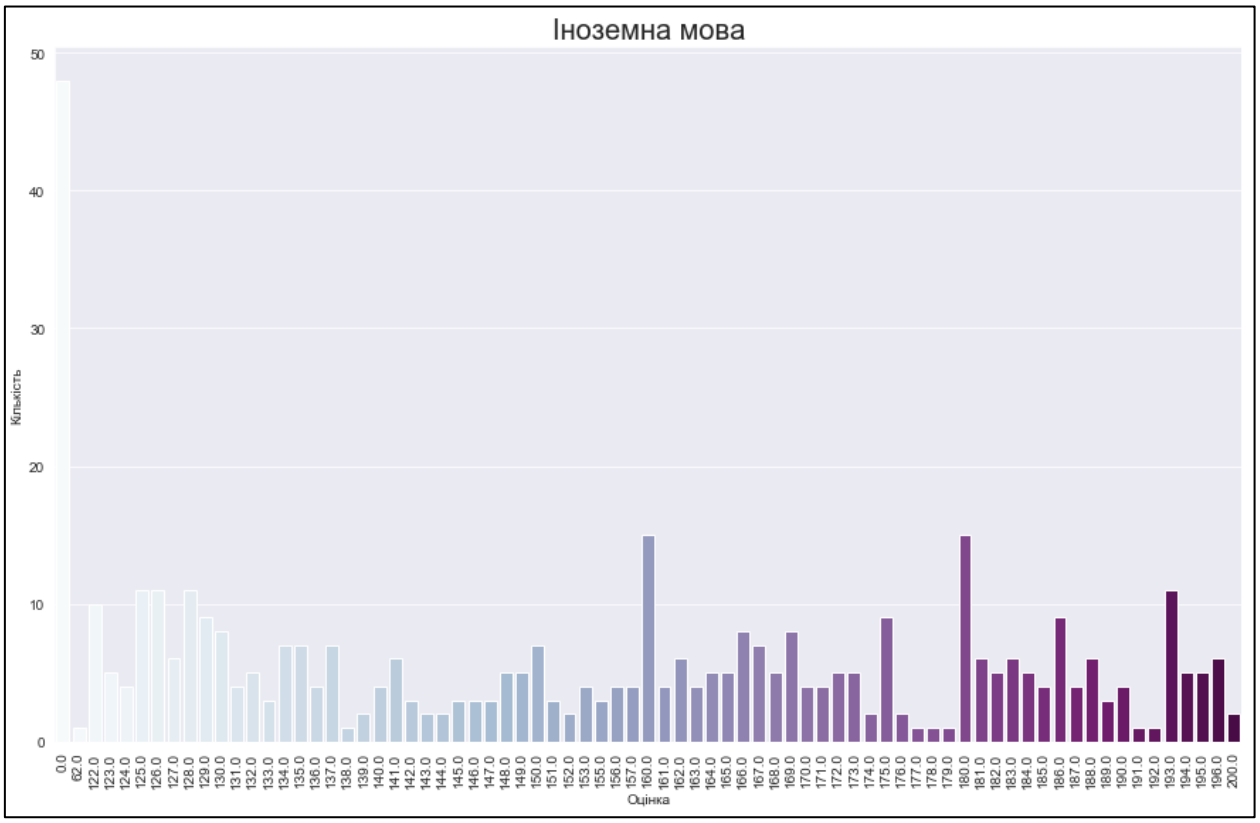

Рисунок 2.40 – Гістограма розподілу дисципліни «Іноземна мова»

Наступним було проаналізовано другий датасет. На рисунку 2.41 продемонстровано кореляційну матрицю його значень.

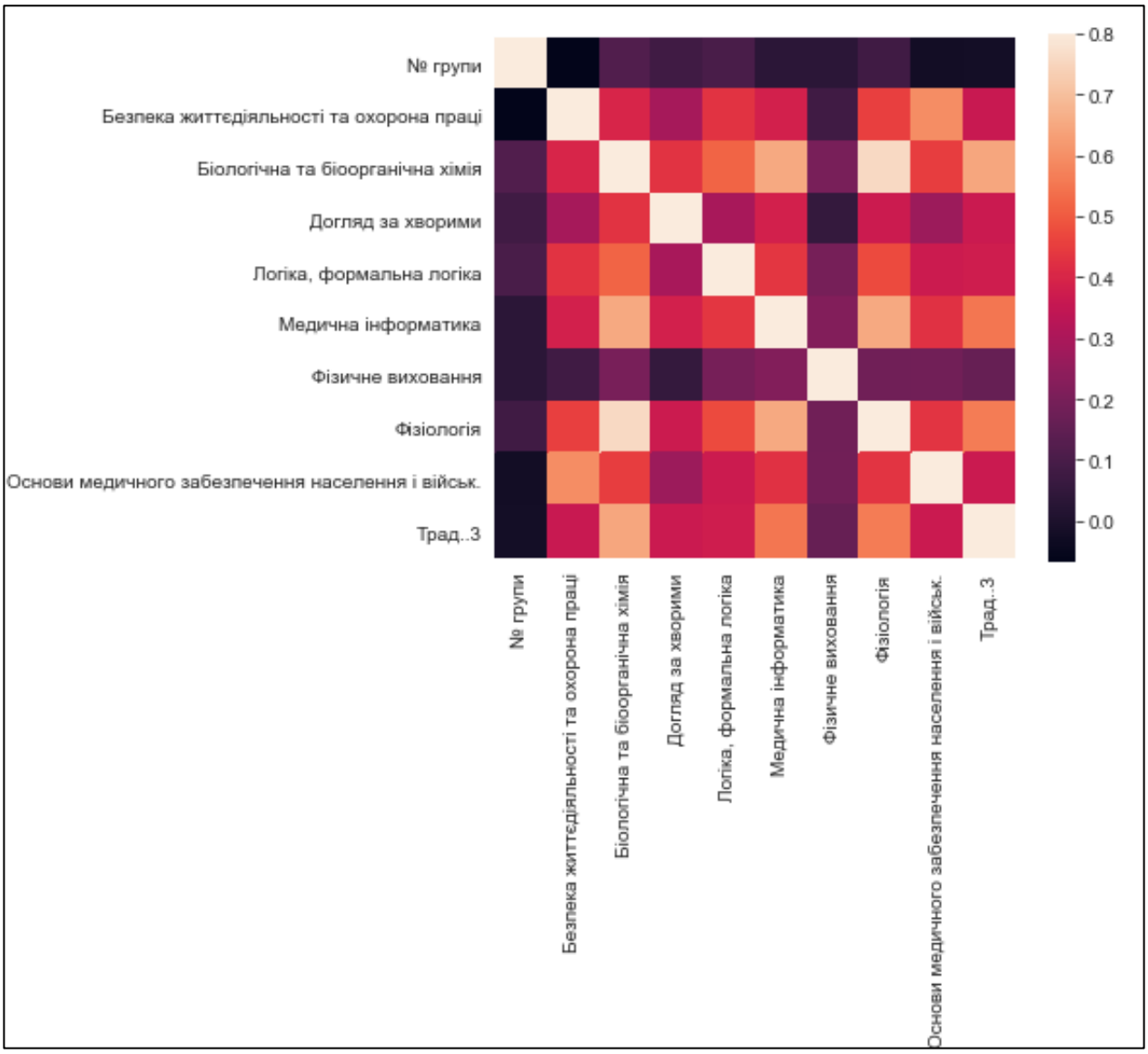

Рисунок 2.41 – Теплова кореляційна карта

На основі цієї карти можна зробити висновок, що існує деяка залежність між колонками «Фізіологія» та «Біологічна та біоорганічна хімія». Найвищу кореляцію з англійською мовою з усіх доступних предметів має «Біологічна та біоорганічна хімія».

Далі було побудовано графіки розподілу значень по усіх дисциплінах. Ці графіки та код для їх побудови продемонстровано на рисунках 2.42-2.43.

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Іноземна мова за професійним спрямуванням')
sns.violinplot(y='Іноземна мова за професійним спрямуванням',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Безпека життєдіяльності та охорона праці')
sns.violinplot(y='Безпека життєдіяльності та охорона праці',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Біологічна та біоорганічна хімія')
sns.violinplot(y='Біологічна та біоорганічна хімія',data=small,palette='spring',linewidth=3)
plt.show()
```

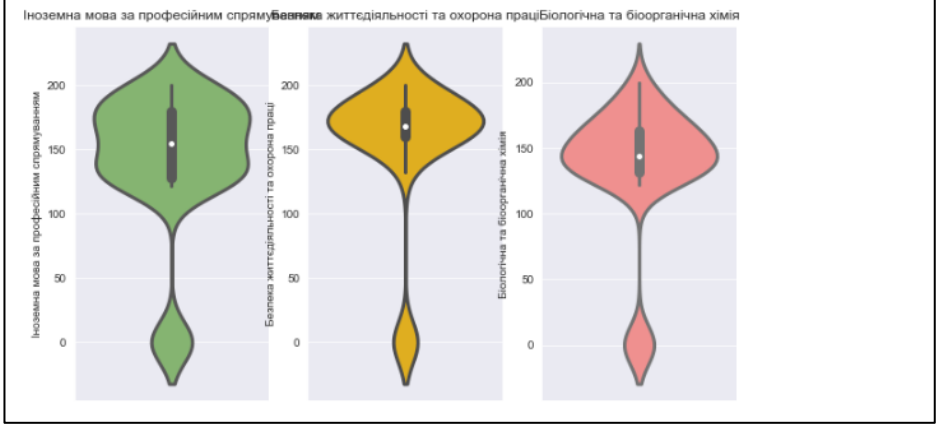

Рисунок 2.42 – Графіки розподілу значень перших трьох дисциплін

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Догляд за хворими')
sns.violinplot(y='Догляд за хворими',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Логіка, формальна логіка')
sns.violinplot(y='Логіка, формальна логіка',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Медична інформатика')
sns.violinplot(y='Медична інформатика',data=small,palette='spring',linewidth=3)
plt.show()
```

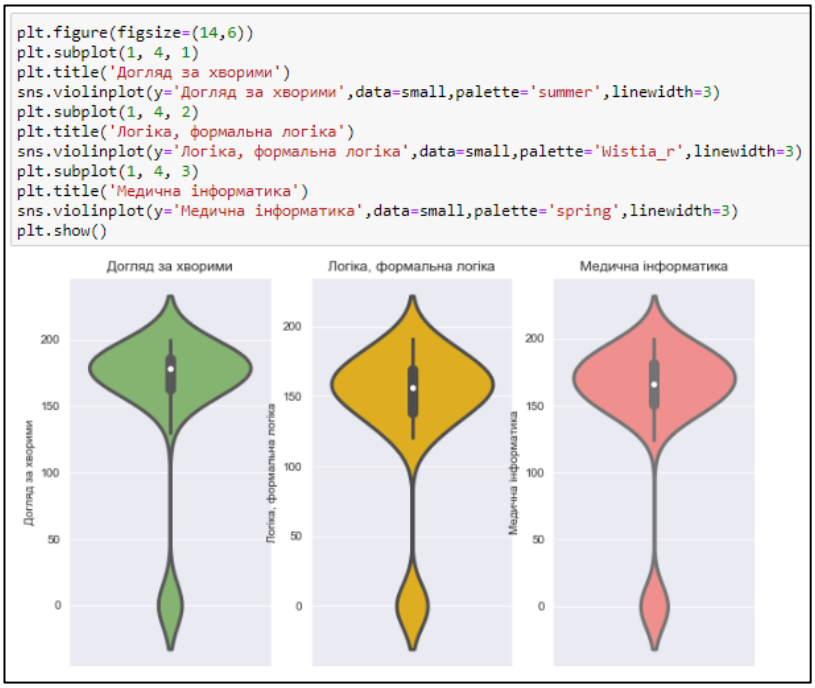

Рисунок 2.43 – Графіки розподілу значень наступних трьох дисциплін

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Фізичне виховання')
sns.violinplot(y='Фізичне виховання',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Фізіологія')
sns.violinplot(y='Фізіологія',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Цивільний захист')
sns.violinplot(y='Цивільний захист. Основи медичного забезпечення населення і військ.'
               ,data=small,palette='spring',linewidth=3)
plt.show()
```

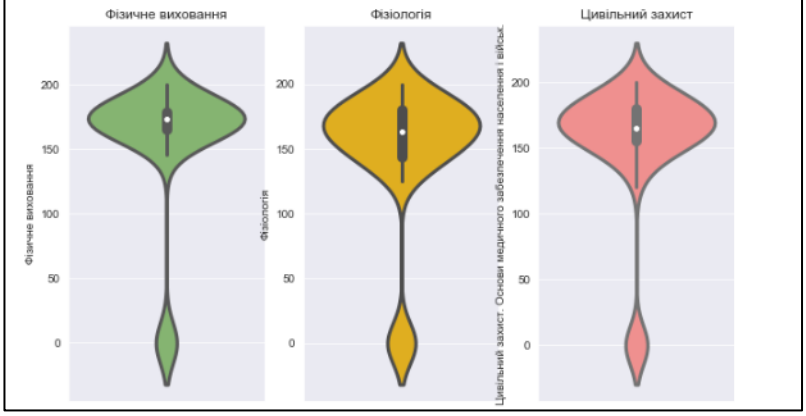

Рисунок 2.44 – Графіки розподілу значень останніх трьох дисциплін

Графік розподілу дисципліни «Іноземна мова за професійним спрямуванням» відмінний від усіх інших, тому не є очевидним зв’язок з будь-якою дисципліною.

Також розглянемо гістограму розподілу дисципліни «Іноземна мова за професійним спрямуванням» (рис. 2.45).

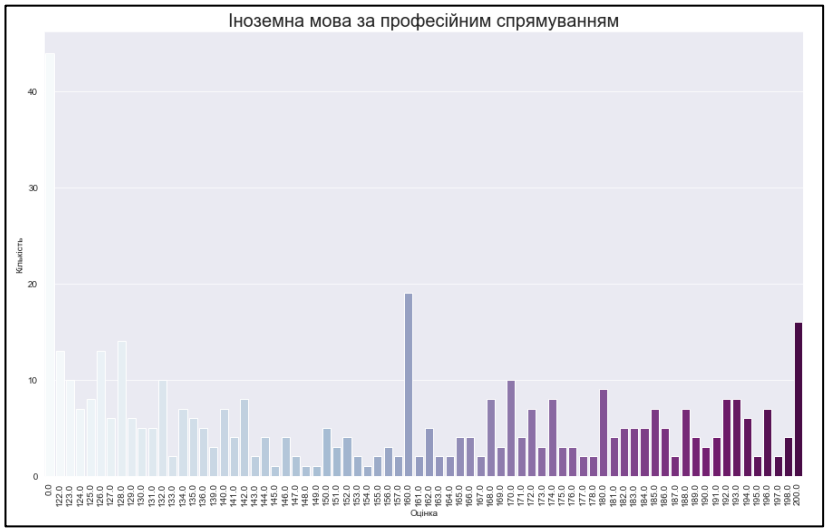

Рисунок 2.45 – Гістограма розподілу дисципліни «Іноземна мова за професійним спрямуванням»

Після цього було проаналізовано датасет з даними за 6 семестр навчання. На рисунку 2.46 продемонстровано кореляційну матрицю його значень.

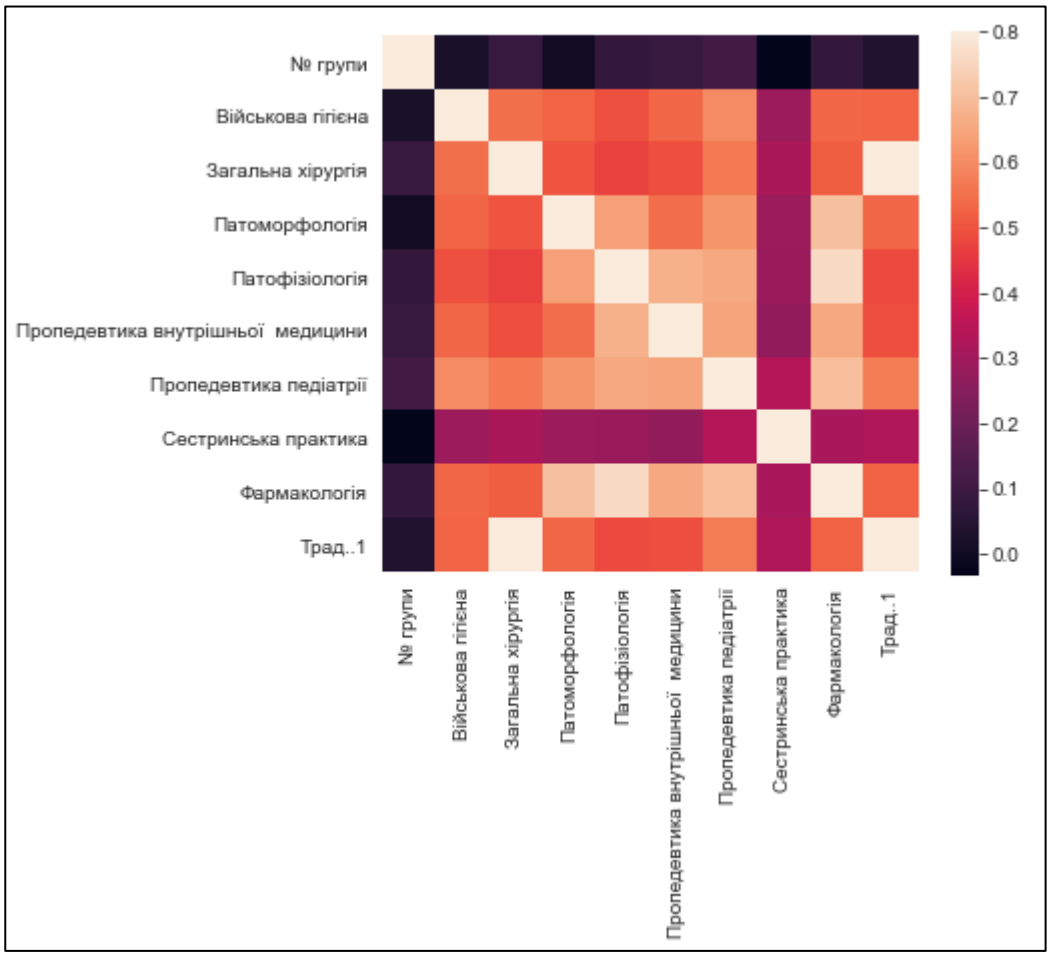

Рисунок 2.46 – Теплова кореляційна карта

На основі цієї карти можна зробити висновок, що існує деяка залежність між колонками «Патофізіологія» та «Фармакологія». Найвищу кореляцію з англійською мовою з усіх доступних предметів має «Патофізіологія».

Далі було побудовано графіки розподілу значень по усіх дисциплінах. Ці графіки та код для їх побудови продемонстровано на рисунках 2.47-2.49.

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Іноземна мова')
sns.violinplot(y='Іноземна мова',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Військова гігієна')
sns.violinplot(y='Військова гігієна',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Загальна хірургія')
sns.violinplot(y='Загальна хірургія',data=small,palette='spring',linewidth=3)
plt.show()
```

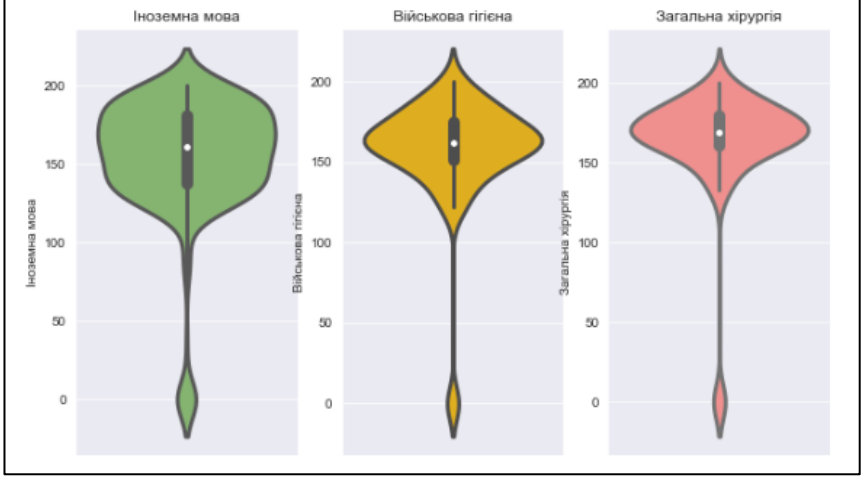

Рисунок 2.47 – Графіки розподілу значень останніх трьох дисциплін

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Патоморфологія')
sns.violinplot(y='Патоморфологія',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Патофізіологія')
sns.violinplot(y='Патофізіологія',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Пропедевтика внутрішньої  медицини')
sns.violinplot(y='Пропедевтика внутрішньої  медицини',data=small,palette='spring',linewidth=3)
plt.show()
```

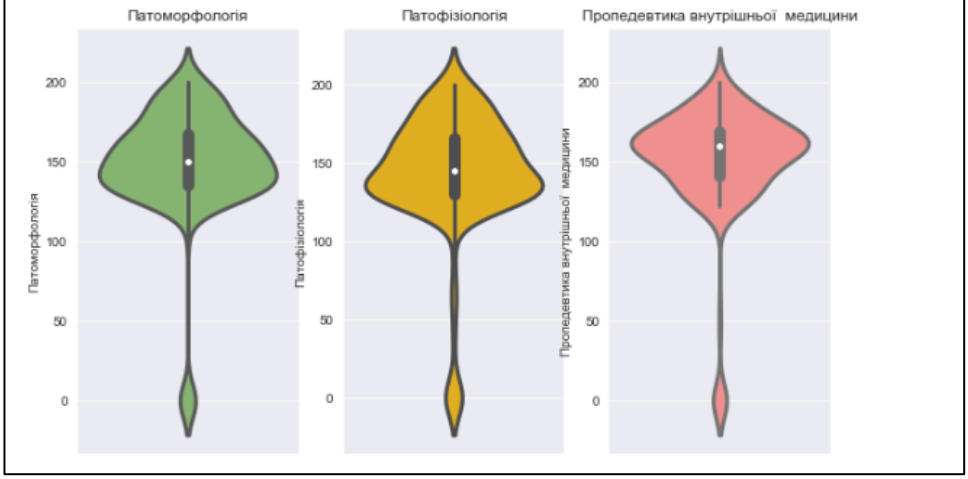

Рисунок 2.48 – Графіки розподілу значень останніх трьох дисциплін

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Пропедевтика педіатрії')
sns.violinplot(y='Пропедевтика педіатрії',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Сестринська практика')
sns.violinplot(y='Сестринська практика',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Фармакологія')
sns.violinplot(y='Фармакологія',data=small,palette='spring',linewidth=3)
plt.show()
```

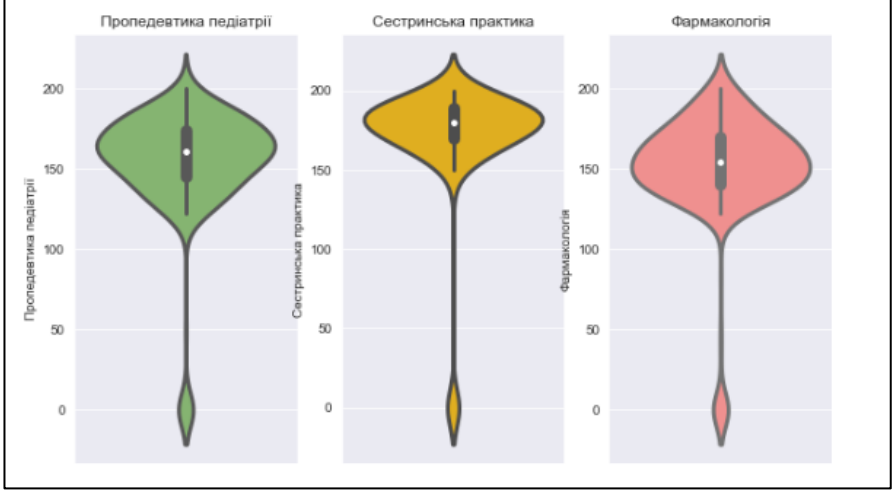

Рисунок 2.49 – Графіки розподілу значень останніх трьох дисциплін

Знову-таки, розподіл дисципліни «Іноземна мова» не схожий на графіки інших дисциплін. Також розглянемо гістограму розподілу дисципліни «Іноземна мова» (рис. 2.50).

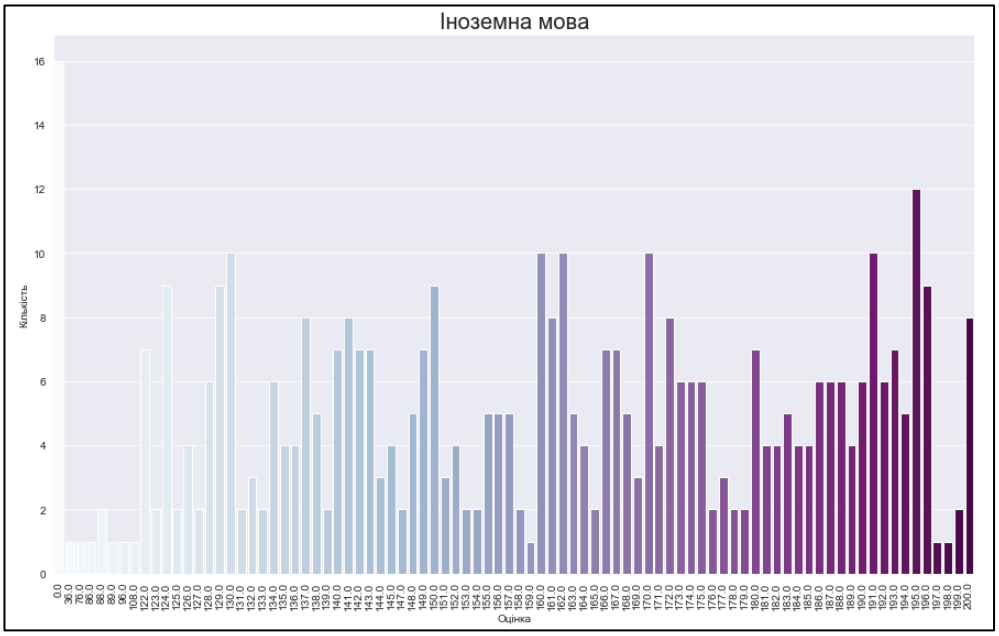

Рисунок 2.50 – Гістограма розподілу дисципліни «Іноземна мова»

Останнім кроком у розвідувальному аналізі була побудова кореляційної матриці датасету з значеннями з дисциплін за 2, 4, 6 семестри навчання та результатами проходження ЄДКІ з англійської мови (рис. 2.51).

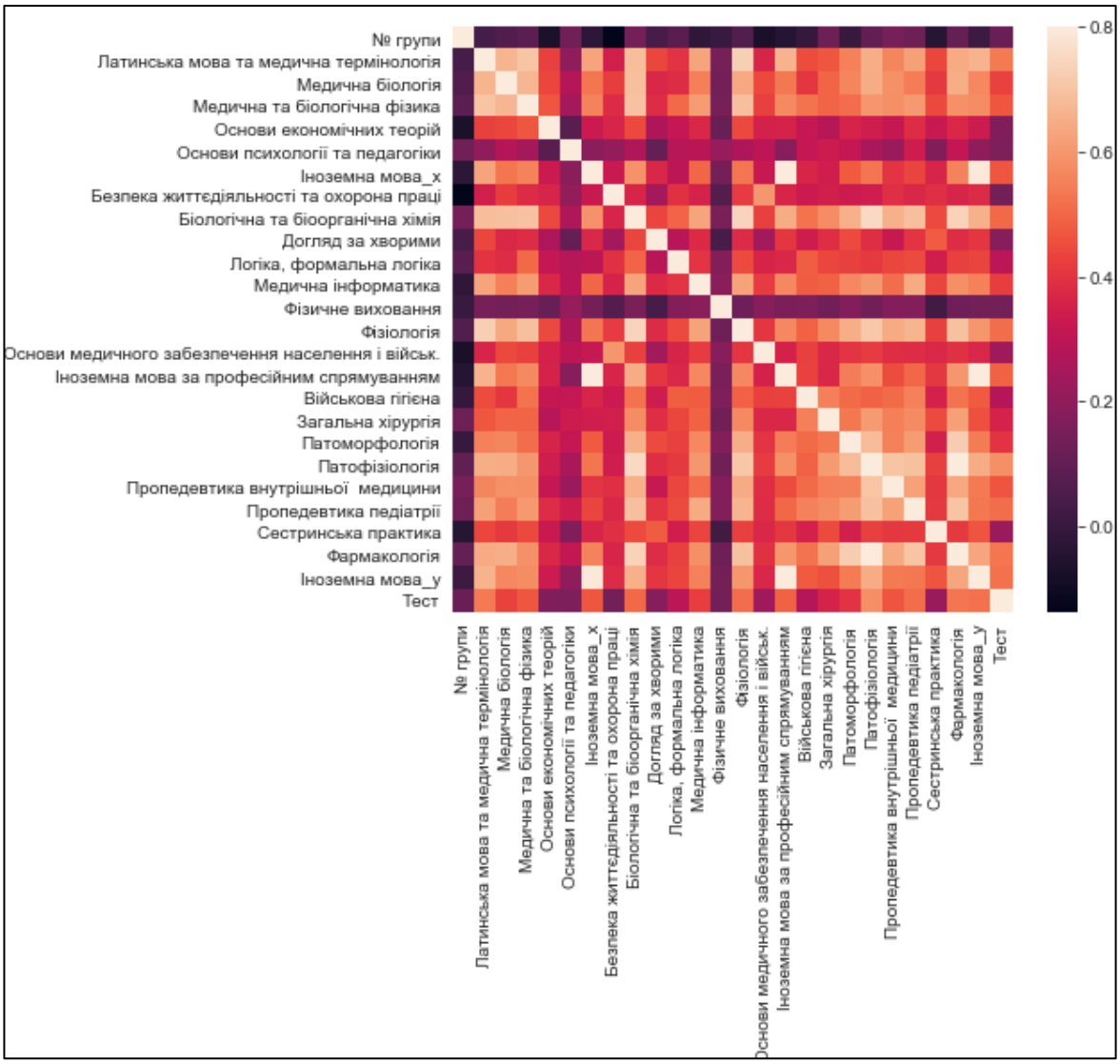

Рисунок 2.51 – Теплова кореляційна карта усіх значень

На ній видно всі ті самі залежності, що і описано вище. Але з’явились нові точки з високим коефіцієнтом кореляції, між предметами «Іноземна мова» на різних семестрах навчання.

2.3 Висновки

У другому розділі розроблено метод аналізу початкових даних, за допомогою якого визначено фактори, що впливають на рівень вивчення іноземної мови студентами Вінницького національного медичного університету ім. М. І. Пирогова.

В наступному розділі пропонується розробити моделі з використанням технології дерева рішень з метою екстракції закономірностей між факторами що впливають на рівень знань іноземної мови студентів закладу вищої освіти.

3 РОЗРОБКА МОДЕЛЕЙ ВИЯВЛЕННЯ ЗАКОНОМІРНОСТЕЙ

3.1 Модель об’єкта

Оскільки даних доступних для моделювання не дуже багато, то було використано декілька варіантів моделювання.

Для кожного з 4 датасетів було окремо побудовано моделі. На рисунку 3.1 продемонстровано дерево рішень для датасету з даними за 2 семестр навчання.

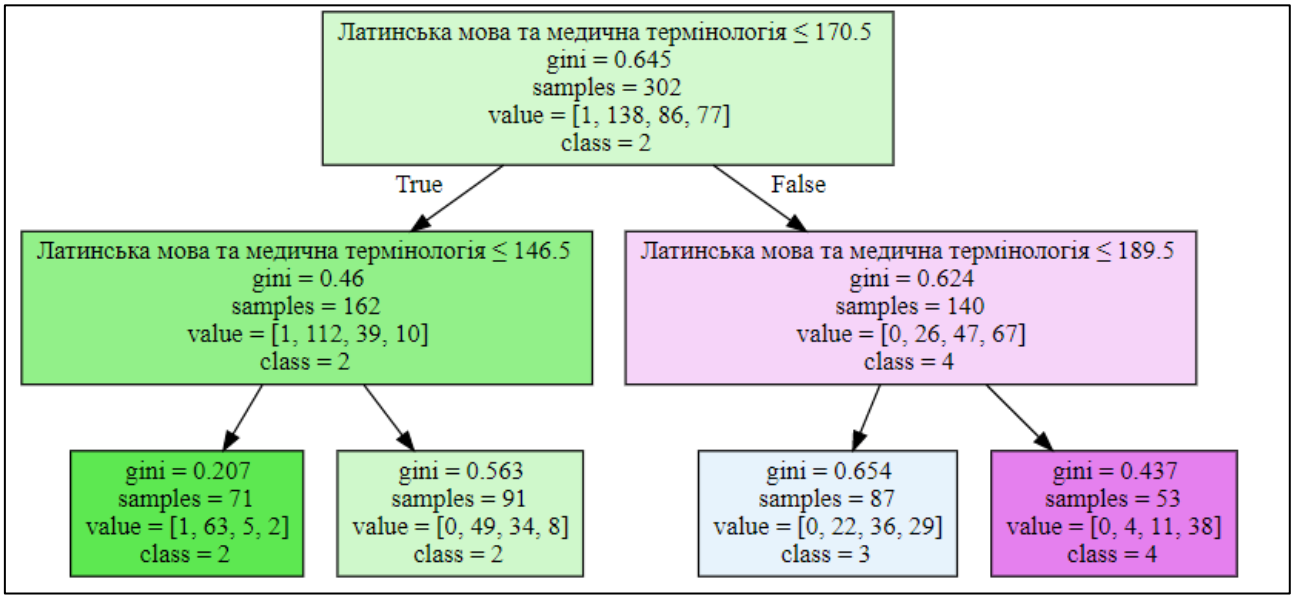

Рисунок 3.1 – Дерево рішень першого датасету

З побудованого дерева рішень видно, що основним параметром його створення є колонка «Латинська мова та медична термінологія».

На рисунку 3.2 показано діаграму важливості ознак lightgbm моделі.

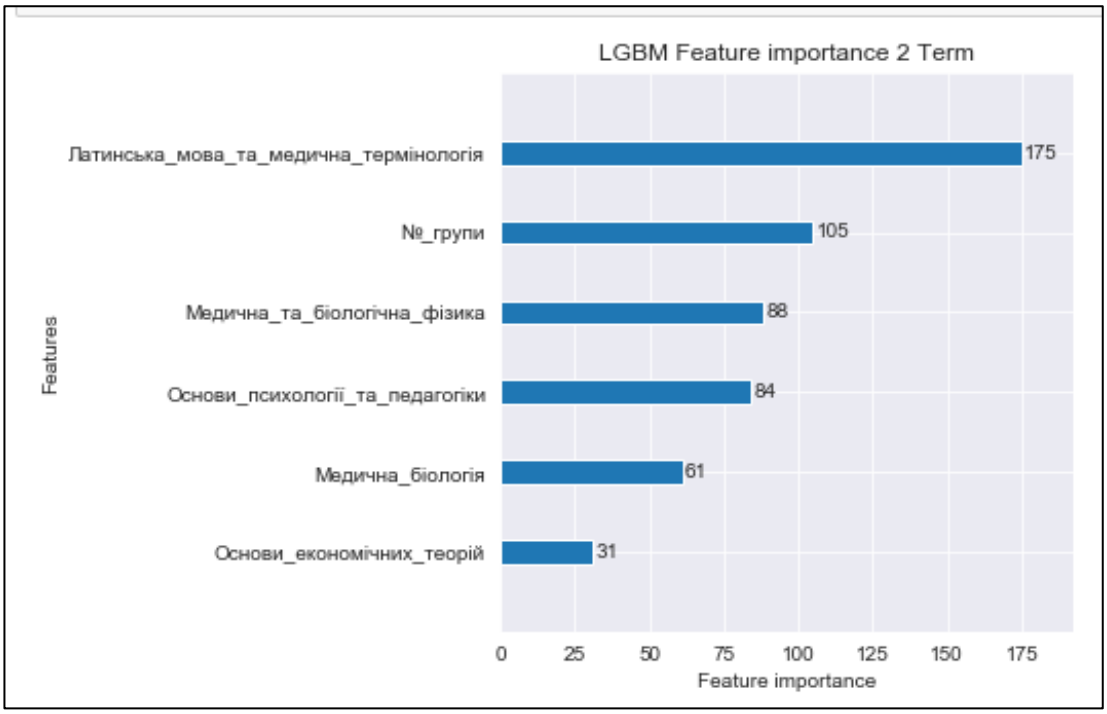

Рисунок 3.2 – Діаграма важливості ознак lightgbm моделі

На рисунку 3.3 показано діаграму важливості ознак xgboost моделі.

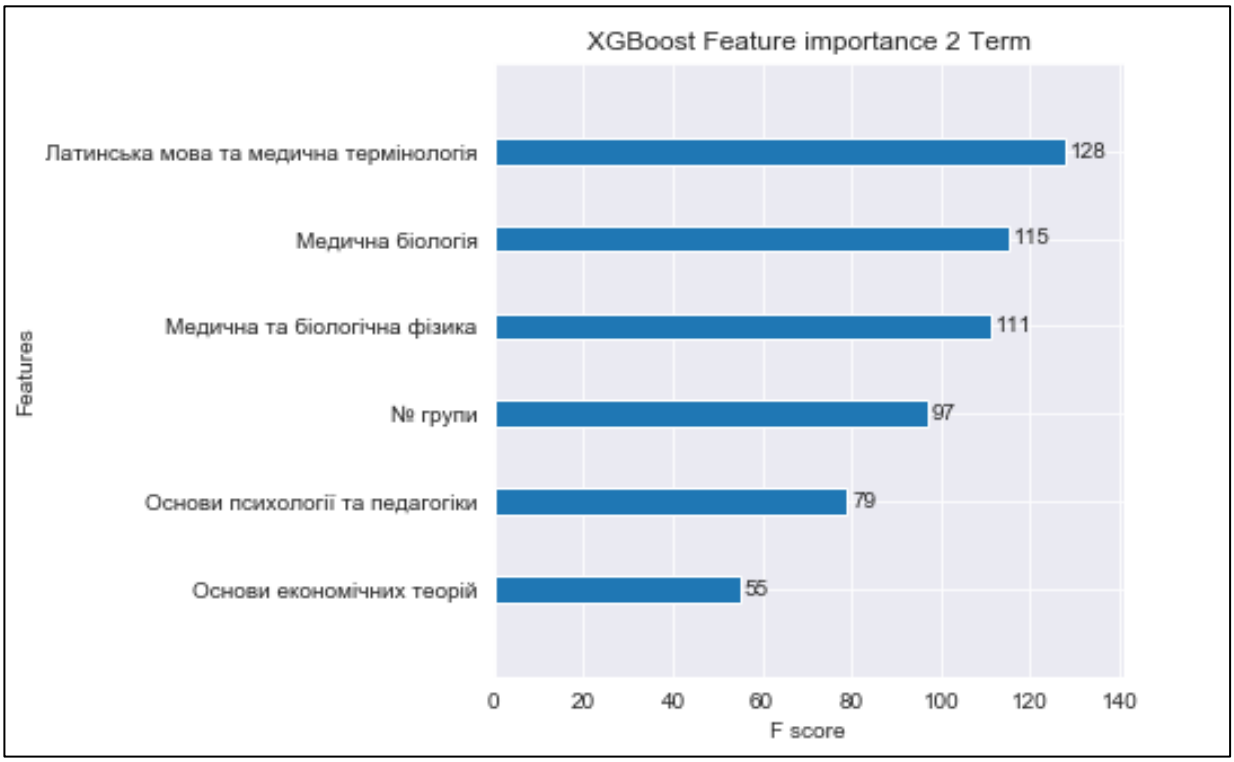

Рисунок 3.3 – Діаграма важливості ознак xgboost моделі

Основною ознакою lightgbm та xgboost моделей стала колонка «Латинська мова та медична термінологія», як і в побудові дерева рішень. На основі цього можна зробити висновок, що «Латинська мова та медична термінологія» є найбільш впливовою на рівень вивчення іноземної мови студентами 2 семестру навчання.

Код побудови моделі логістичної регресії, матриці невідповідностей та сама матриця зображено на рисунках 3.4-3.6 відповідно.

```
from sklearn.linear_model import  LogisticRegression

# creating a model
model = LogisticRegression()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the test set results
y_pred = model.predict(x_test)
```

Рисунок 3.4 – Код побудови моделі логістичної регресії

```
from sklearn.metrics import confusion_matrix

# creating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# printing the confusion matrix
plt.rcParams['figure.figsize'] = (8, 8)
sns.heatmap(cm, annot = True, cmap = 'Greens')
plt.title('Confusion Matrix for Logistic Regression, 2 Term', fontweight = 30, fontsize = 20)
plt.show()
```

Рисунок 3.5 – Код побудови матриці невідповідностей

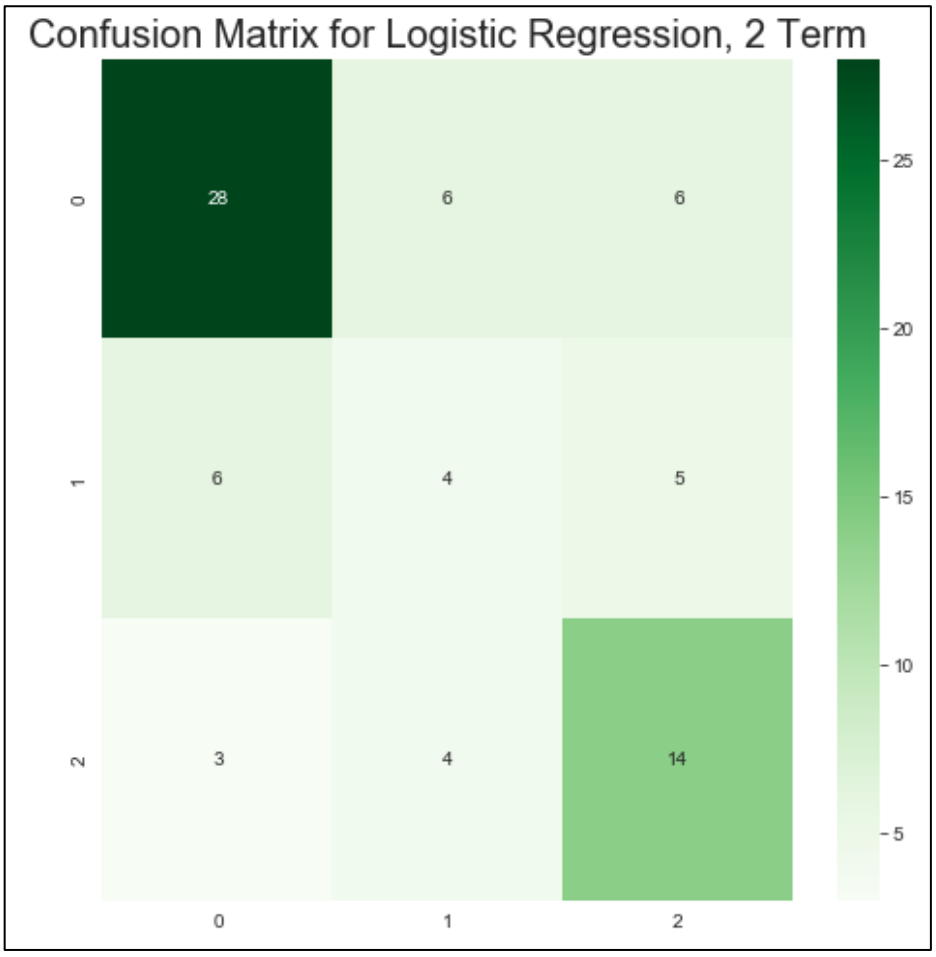

Рисунок 3.6 – Матриця невідповідностей

Точність моделі – 0.6, що є непоганим результатом. Найкраще зпрогонозованими є 1 та 3 класи.

Код побудови моделі Random Forest, матриці невідповідностей та сама матриця зображено на рисунках 3.7-3.9 відповідно.

```
from sklearn.ensemble import RandomForestClassifier

# creating a model
model = RandomForestClassifier()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the x-test results
y_pred = model.predict(x_test)
```

Рисунок 3.7 – Код побудови моделі Random Forest

```
from sklearn.metrics import confusion_matrix

# creating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# printing the confusion matrix
plt.rcParams['figure.figsize'] = (8, 8)
sns.heatmap(cm, annot = True, cmap = 'Reds')
plt.title('Confusion Matrix for Random Forest, 2 Term', fontweight = 30, fontsize = 20)
plt.show()
```

Рисунок 3.8 – Код побудови матриці невідповідностей

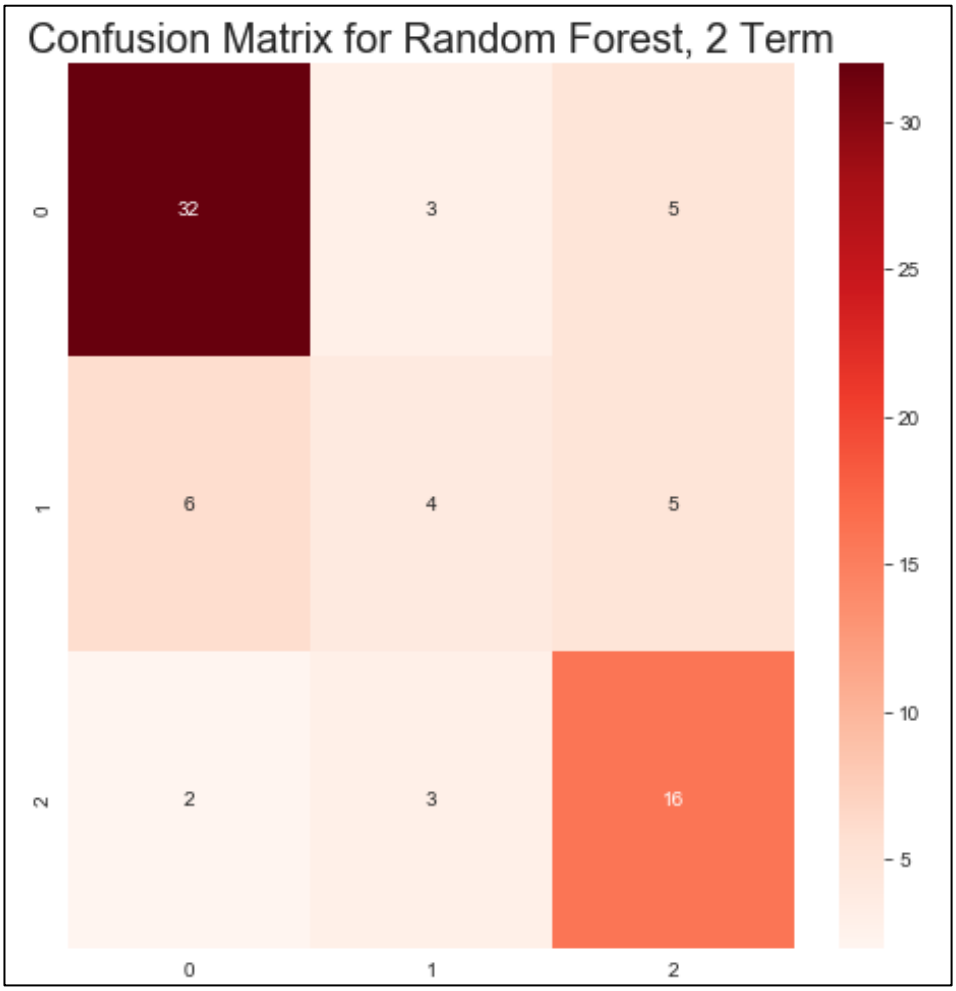

Рисунок 3.9 – Матриця невідповідностей

Точність моделі – 0.7, найкраще зпрогонозованими є 1 та 3 класи.

На рисунку 3.10 продемонстровано дерево рішень для датасету з даними за 4 семестр навчання.

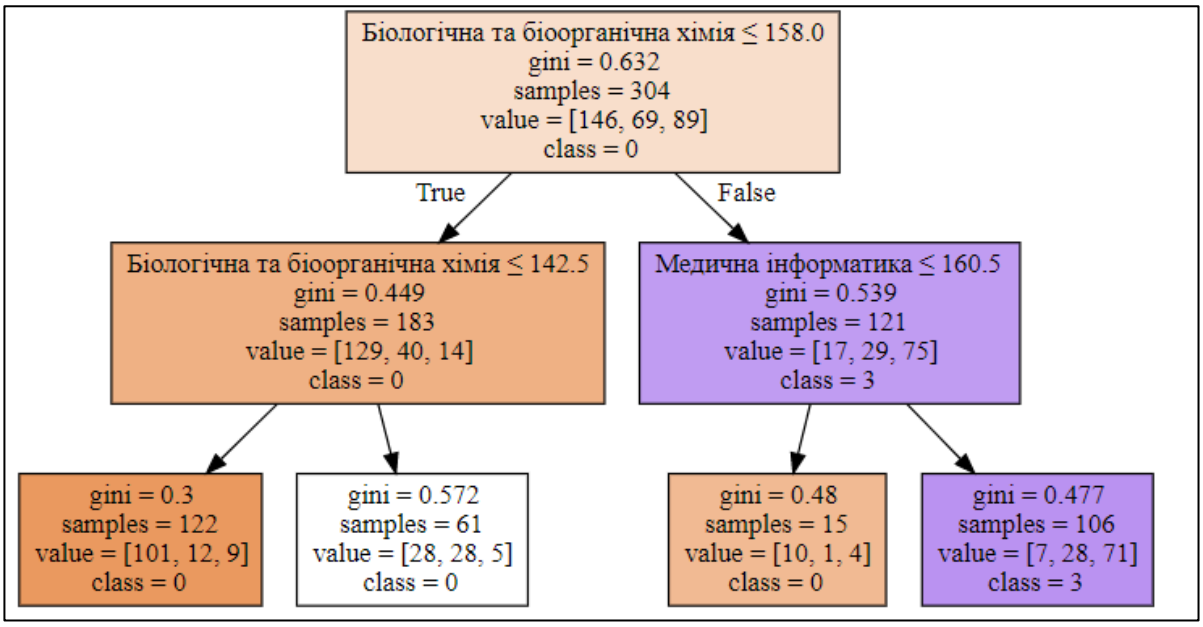

Рисунок 3.10 – Дерево рішень другого датасету

З побудованого дерева рішень видно, що основними параметрами його створення є колонки «Біологічна та біоорганічна хімія» та «Медична інформатика».

На рисунку 3.11 показано діаграму важливості ознак lightgbm моделі.

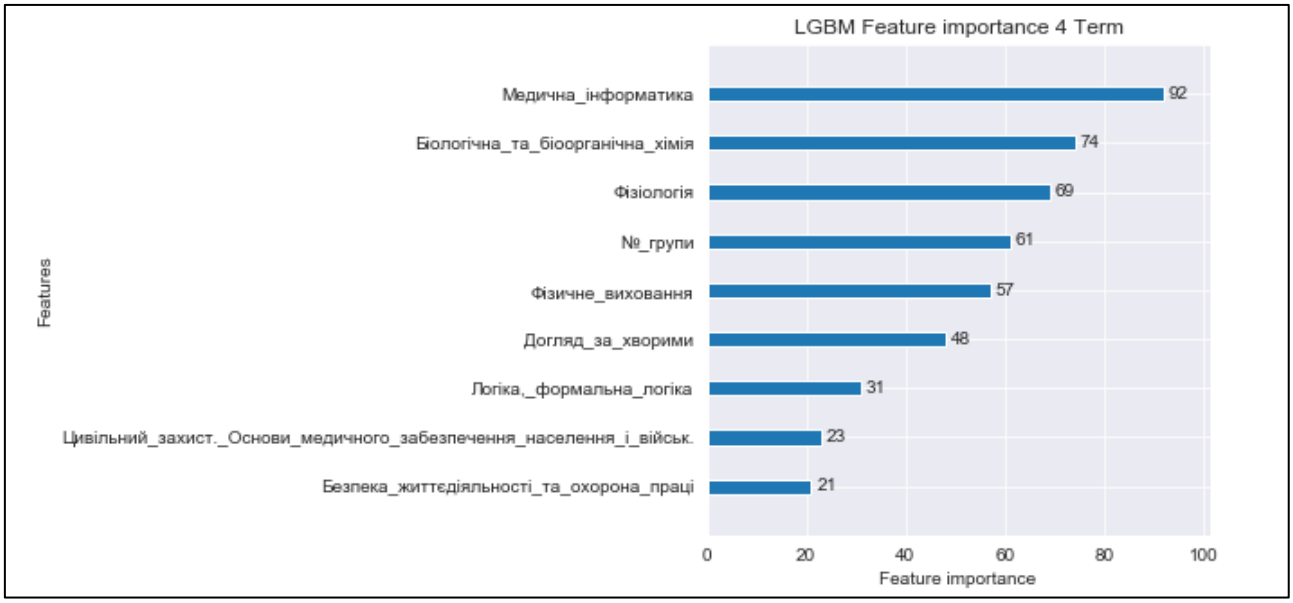

Рисунок 3.11 – Діаграма важливості ознак lightgbm моделі

На рисунку 3.12 показано діаграму важливості ознак xgboost моделі.

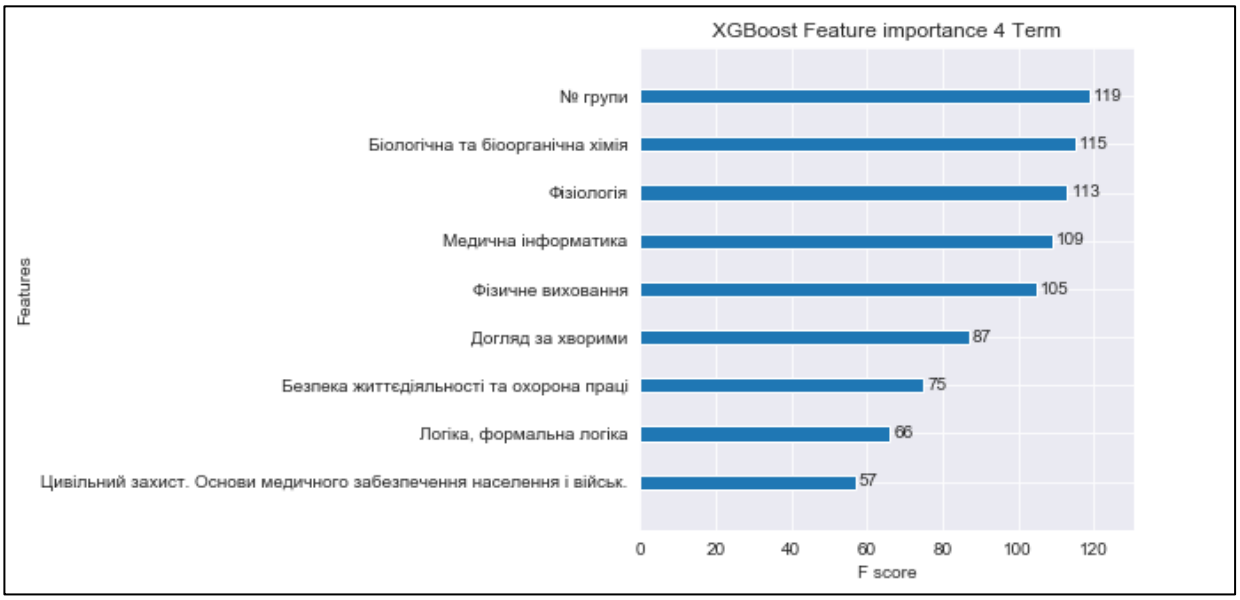

Рисунок 3.12 – Діаграма важливості ознак xgboost моделі

Основною ознакою lightgbm стала колонка «Медична інформатика», як і в побудові дерева рішень. Але на діаграмі ознак xgboost вона зайняла 5 місце, проте перші 5 ознак мають дуже близькі значення важливості. На основі цього можна зробити висновок, що «Медична інформатика» є найбільш впливовою на рівень вивчення іноземної мови студентами 4 семестру навчання.

Код побудови моделі логістичної регресії, матриці невідповідностей та сама матриця зображено на рисунках 3.13-3.15 відповідно.

```
from sklearn.linear_model import  LogisticRegression

# creating a model
model = LogisticRegression()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the test set results
y_pred = model.predict(x_test)
```

Рисунок 3.13 – Код побудови моделі логістичної регресії

```python
from sklearn.metrics import confusion_matrix

# creating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# printing the confusion matrix
plt.rcParams['figure.figsize'] = (8, 8)
sns.heatmap(cm, annot = True, cmap = 'Greens')
plt.title('Confusion Matrix for Logistic Regression, 4 Term', fontweight = 30, fontsize = 20)
plt.show()
```

Рисунок 3.14 – Код побудови матриці невідповідностей

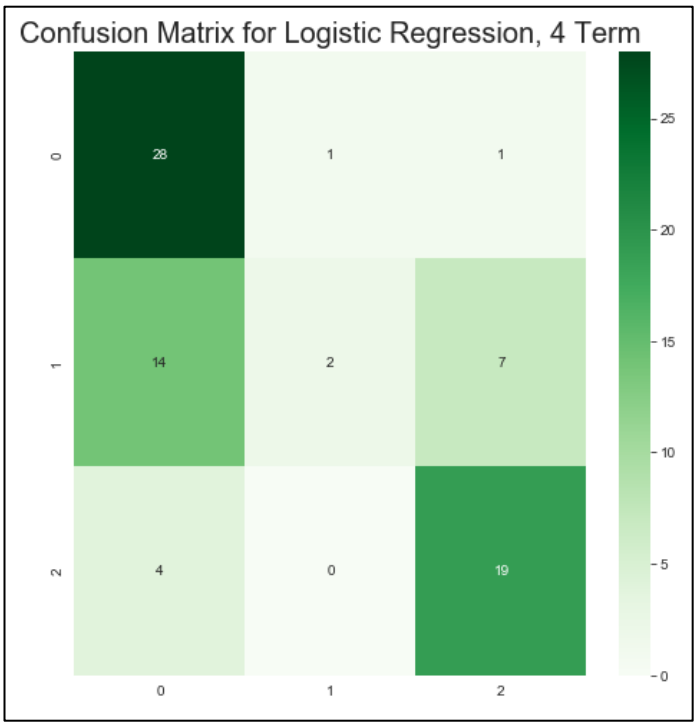

Рисунок 3.15 – Матриця невідповідностей

Точність моделі – 0.64, найкраще зпрогонозованими є 1 та 3 класи.

Код побудови моделі Random Forest, матриці невідповідностей та сама матриця зображено на рисунках 3.16-3.18 відповідно.

```python
from sklearn.ensemble import RandomForestClassifier

# creating a model
model = RandomForestClassifier()

# fitting the training data to the model
model.fit(x_train, y_train)

# predicting the x-test results
y_pred = model.predict(x_test)
```

Рисунок 3.16 – Код побудови моделі Random Forest

```
from sklearn.metrics import confusion_matrix

# creating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# printing the confusion matrix
plt.rcParams['figure.figsize'] = (8, 8)
sns.heatmap(cm, annot = True, cmap = 'Reds')
plt.title('Confusion Matrix for Random Forest, 4 Term', fontweight = 30, fontsize = 20)
plt.show()
```

Рисунок 3.17 – Код побудови матриці невідповідностей

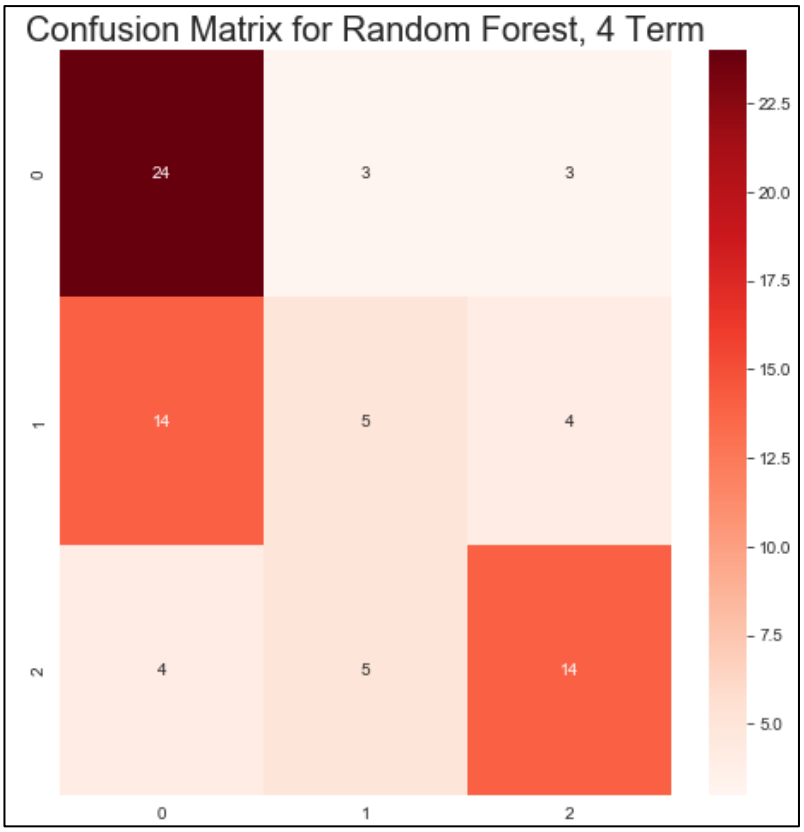

Рисунок 3.18 – Матриця невідповідностей

Точність моделі – 0.60, найкраще зпрогонозованими є 1 та 3 класи.

На рисунку 3.19 продемонстровано дерево рішень для датасету з даними за 6 семестр навчання.

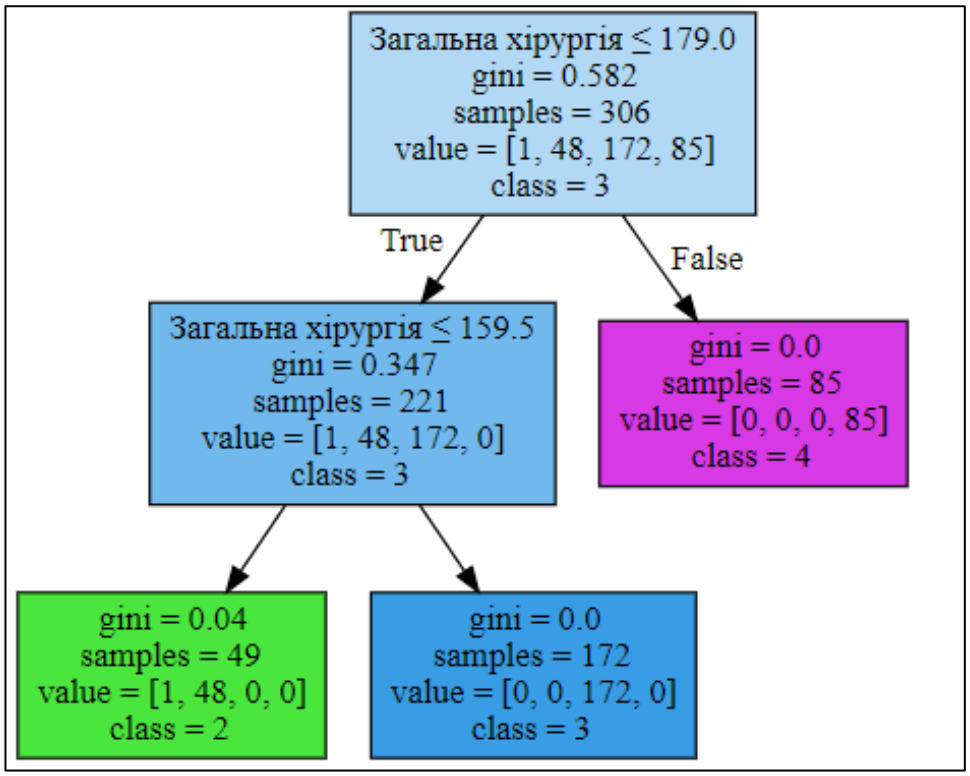

Рисунок 3.19 – Дерево рішень третього датасету

З побудованого дерева рішень видно, що основним параметром його створення є колонка «Загальна хірургія».

На рисунку 3.20 показано діаграму важливості ознак lightgbm моделі.

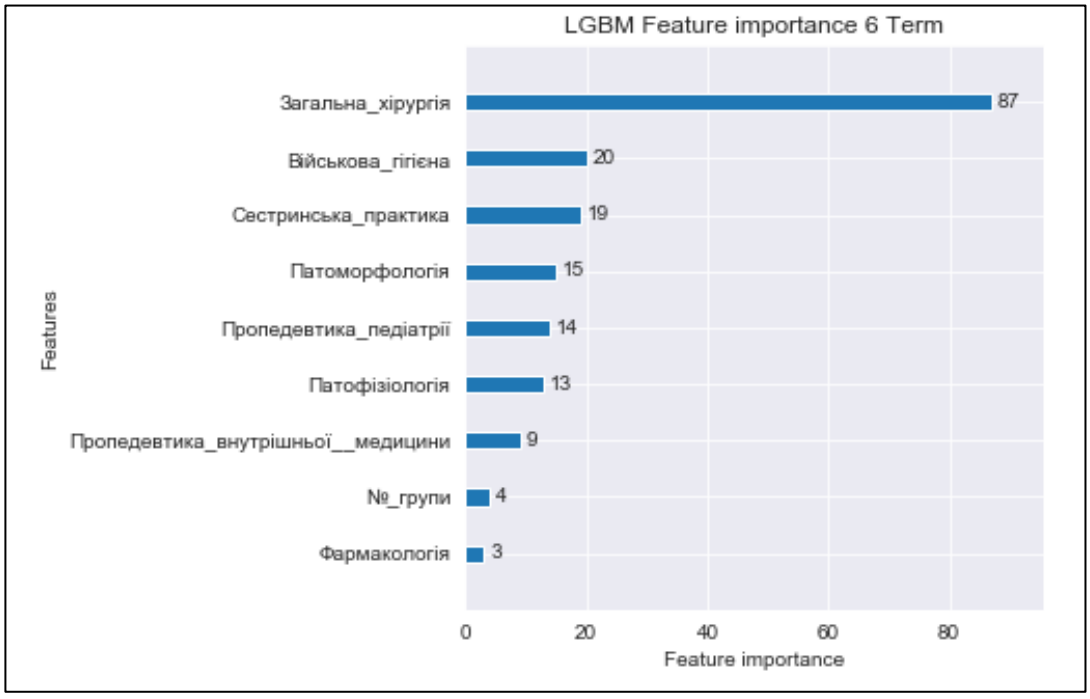

Рисунок 3.20 – Діаграма важливості ознак lightgbm моделі

На рисунку 3.21 показано діаграму важливості ознак xgboost моделі.

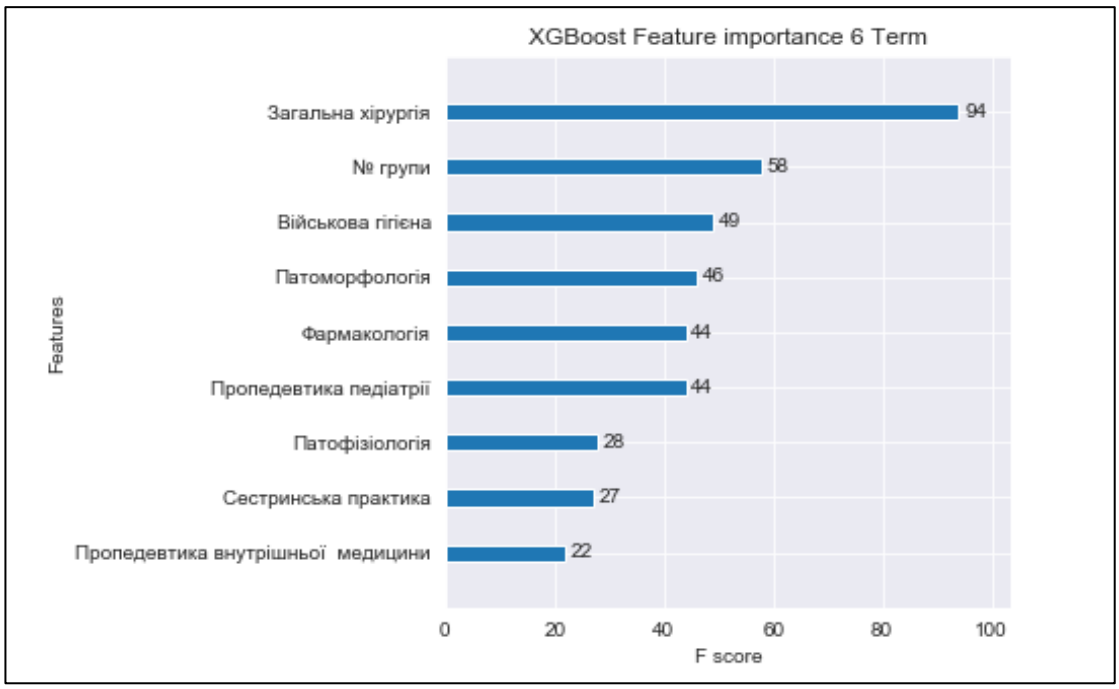

Рисунок 3.21 – Діаграма важливості ознак xgboost моделі

Основною ознакою lightgbm та xgboost моделей однозначно стала колонка «Загальна хірургія», як і в побудові дерева рішень. На основі цього можна зробити висновок, що «Загальна хірургія» є найбільш впливовою на рівень вивчення іноземної мови студентами 6 семестру навчання.

Код побудови моделі логістичної регресії, матриці невідповідностей та сама матриця зображено на рисунках 3.23-3.25 відповідно.

```
from sklearn.linear_model import  LogisticRegression

# creating a model
model = LogisticRegression()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the test set results
y_pred = model.predict(x_test)
```

Рисунок 3.23 – Код побудови моделі логістичної регресії

```python
from sklearn.metrics import confusion_matrix

# creating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# printing the confusion matrix
plt.rcParams['figure.figsize'] = (8, 8)
sns.heatmap(cm, annot = True, cmap = 'Greens')
plt.title('Confusion Matrix for Logistic Regression, 6 Term', fontweight = 30, fontsize = 20)
plt.show()
```

Рисунок 3.24 – Код побудови матриці невідповідностей

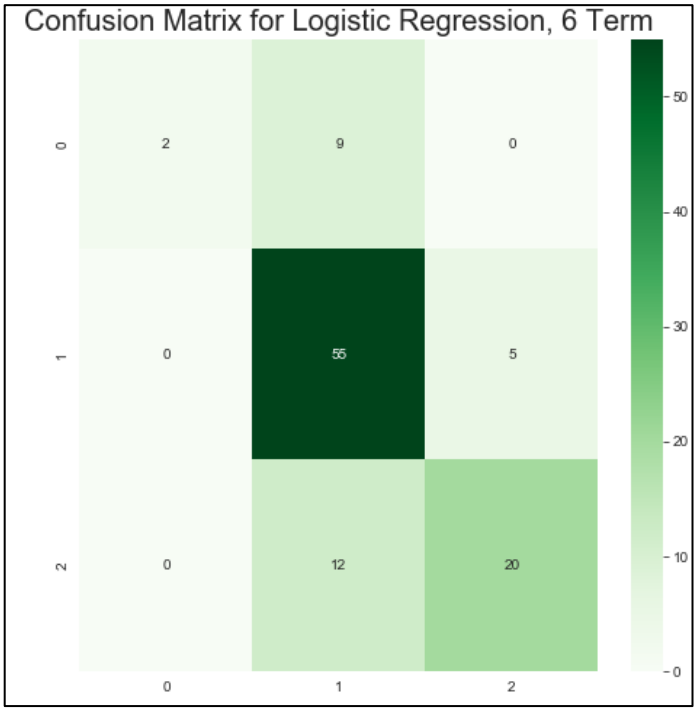

Рисунок 3.25 – Матриця невідповідностей

Точність моделі – 0.75, модель є адекватною. Найкраще зпрогонозованими є 1 та 2 класи.

Код побудови моделі Random Forest, матриці невідповідностей та сама матриця зображено на рисунках 3.26-3.28 відповідно.

```python
from sklearn.ensemble import RandomForestClassifier

# creating a model
model = RandomForestClassifier()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the x-test results
y_pred = model.predict(x_test)
```

Рисунок 3.26 – Код побудови моделі Random Forest

```python
from sklearn.metrics import confusion_matrix

# creating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# printing the confusion matrix
plt.rcParams['figure.figsize'] = (8, 8)
sns.heatmap(cm, annot = True, cmap = 'Reds')
plt.title('Confusion Matrix for Random Forest, 6 Term', fontweight = 30, fontsize = 20)
plt.show()
```

Рисунок 3.27 – Код побудови матриці невідповідностей

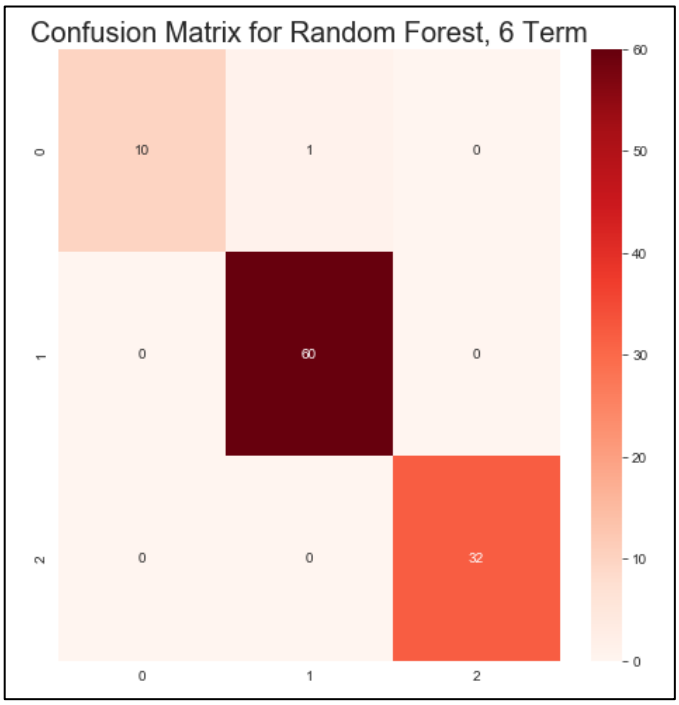

Рисунок 3.28 – Матриця невідповідностей

Точність моделі – 0.99, модель майже ідеально прогнозує усі значення.

Було побудовано моделі для 4 датасету. Код для побудови дерева рішень продемонстровано на рисунку 3.29 .

```
import graphviz
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier, export_graphviz

# Tuning the DecisionTreeClassifier by the GridSearchCV
parameters = {'max_depth' : np.arange(2, 9, dtype=int),
              'min_samples_leaf' :  np.arange(1, 3, dtype=int)}
classifier = DecisionTreeClassifier(random_state=1000)
model = GridSearchCV(estimator=classifier, param_grid=parameters, scoring='accuracy', cv=10, n_jobs=-1)
model.fit(x_train, y_train)
best_parameters = model.best_params_
print(best_parameters)
```

```
{'max_depth': 2, 'min_samples_leaf': 1}
```

```
model=DecisionTreeClassifier(max_depth = best_parameters['max_depth'],
                             random_state = 1118)
model.fit(x_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=2, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=1118, splitter='best')
```

```
# plot tree
dot_data = export_graphviz(model, out_file=None, feature_names=x_train.columns, class_names=['0','2','3','4','5'],
                           filled=True, rounded=False,special_characters=True)
graph = graphviz.Source(dot_data)
graph
```

Рисунок 3.29 – Код для побудови дерева рішень

Побудоване дерево рішень представлено в Додатку Г.

Код побудови моделі lightgbm та діаграми важливості зображено на рисунках 3.30 та 3.31 відповідно.

```
Xtrain, Xval, Ztrain, Zval = train_test_split(x, y, test_size=0.2, random_state=0)
train_set = lgbm.Dataset(Xtrain, Ztrain, silent=False)
valid_set = lgbm.Dataset(Xval, Zval, silent=False)
```

```
params = {
        'boosting_type':'gbdt',
        'objective': 'regression',
        'num_leaves': 31,
        'learning_rate': 0.05,
        'max_depth': -1,
        'subsample': 0.8,
        'bagging_fraction' : 1,
        'max_bin' : 5000 ,
        'bagging_freq': 20,
        'colsample_bytree': 0.6,
        'metric': 'rmse',
        'min_split_gain': 0.5,
        'min_child_weight': 1,
        'min_child_samples': 10,
        'scale_pos_weight':1,
        'zero_as_missing': True,
        'seed':0,
    }

modelL = lgbm.train(params, train_set = train_set, num_boost_round=1000,
                    early_stopping_rounds=50, verbose_eval=10, valid_sets=valid_set)
```

Рисунок 3.30 – Код побудови моделі lightgbm

```
fig =  plt.figure(figsize = (15,15))
axes = fig.add_subplot(111)
lgbm.plot_importance(modelL,ax = axes,height = 0.5)
plt.show();plt.close()
```

Рисунок 3.31 – Код побудови діаграми важливості lightgbm

Діаграма важливості ознак моделі lightgbm, зображена на рисунку 3.32.

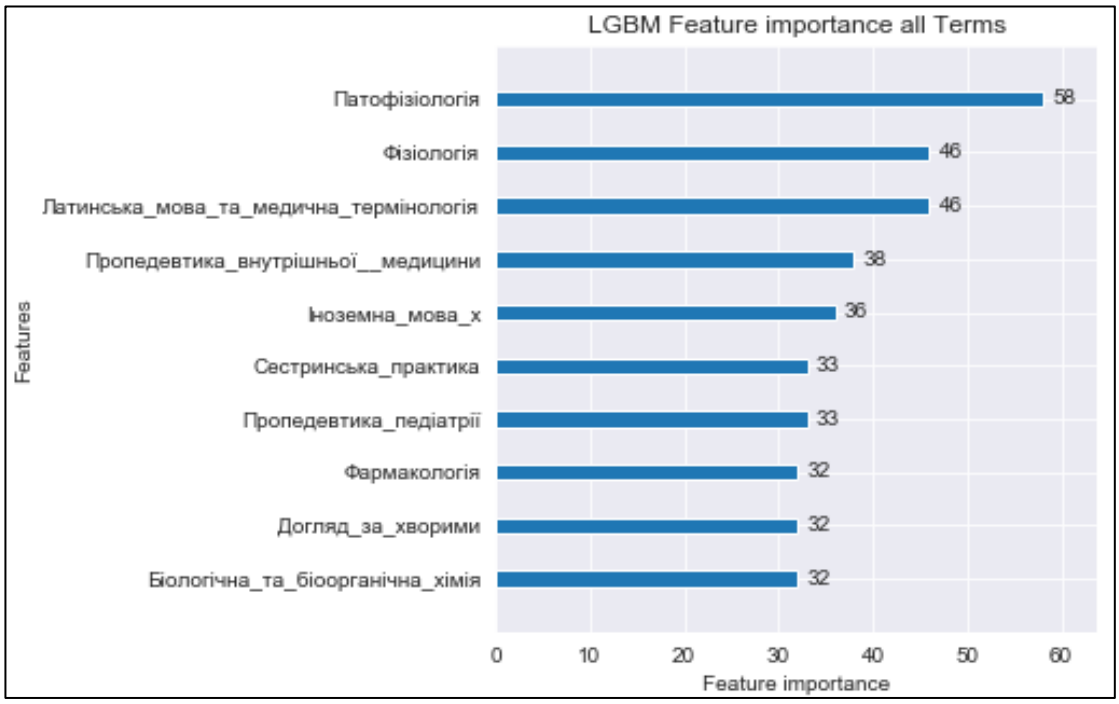

Рисунок 3.32 – Діаграма важливості ознак моделі lightgbm

Код побудови моделі lightgbm та діаграми важливості зображено на рисунках 3.33 та 3.34 відповідно.

```
parms = {'max_depth':8, #maximum depth of a tree
         'objective':'reg:linear',
         'subsample':0.8,#SGD will use this percentage of data
         'lambda '  :4, #L2 regularization term,>1 more conservative
         'colsample_bytree ':0.9,
         'colsample_bylevel':1,
         'min_child_weight': 10}
modelx = xgb.train(parms, data_tr, num_boost_round=200, evals = evallist,
                   early_stopping_rounds=30, maximize=False,
                   verbose_eval=10)

print('score = %1.5f, n_boost_round =%d.'%(modelx.best_score,modelx.best_iteration))
```

Рисунок 3.33 – Код побудови моделі xgboost

```
fig =  plt.figure(figsize = (15,15))
axes = fig.add_subplot(111)
xgb.plot_importance(modelx,ax = axes,height = 0.5)
plt.show();plt.close()
```

Рисунок 3.34 – Код побудови діаграми важливості xgboost

Діаграма важливості ознак моделі xgboost, продемонстрована на рисунку 3.35.

Рисунок 3.35 – Діаграма важливості ознак моделі xgboost

Також, було побудовано матриці невідповідності для моделей логістичної регресії (рис. 3.36) та Random forest (рис 3.37).

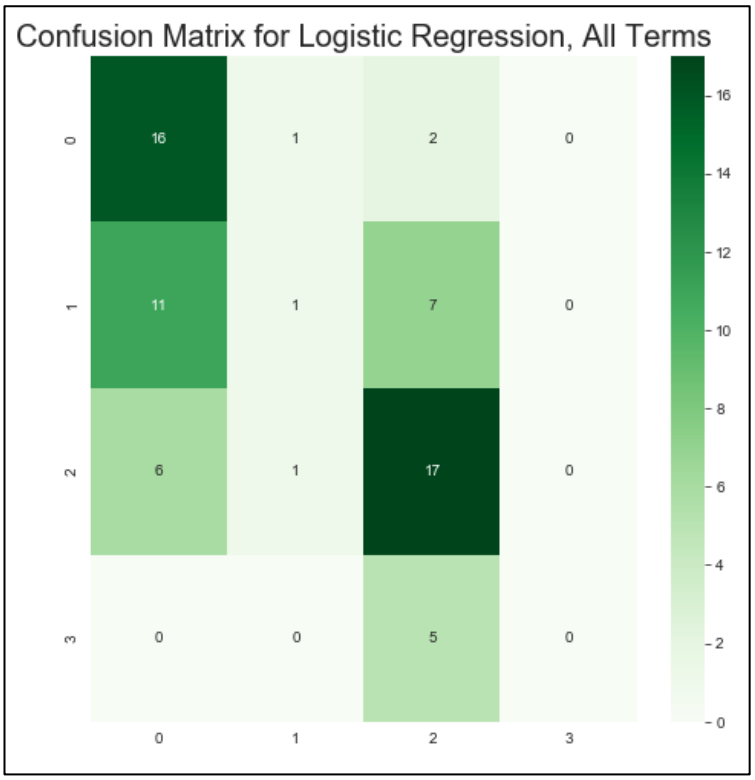

Рисунок 3.36 – Матриця невідповідності логістичної регресії

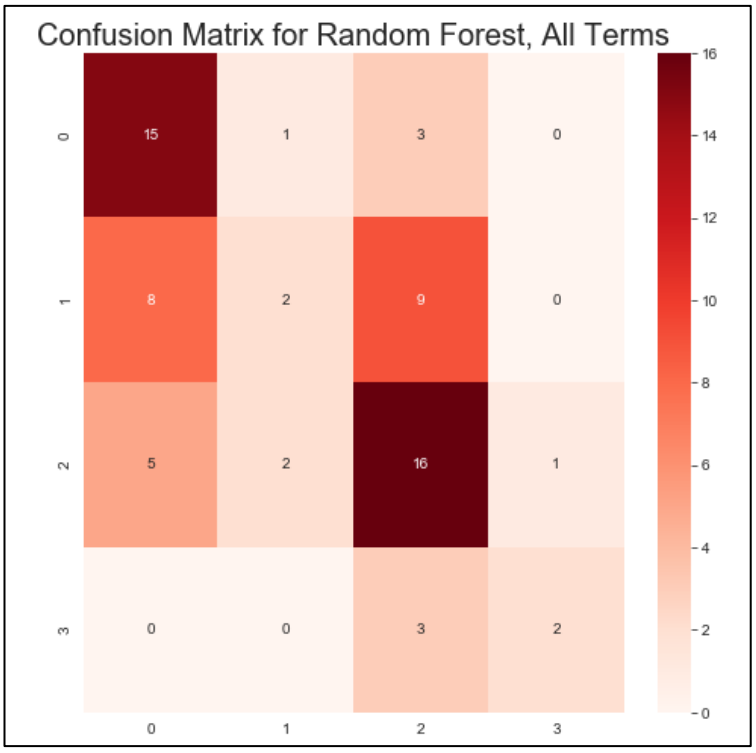

Рисунок 3.37 – Матриця невідповідності Random forest

Точність обох моделей – 0.5, що є доволі низьким показником. Найкраще зпрогонозованими є 1 та 3 класи.

Точність використаних моделей машинного навчання продемонстрована на рисунку 3.38.

	2 семестр	4 семестр	6 семестр	ЄДКІ
LGBM	0.70	0.65	0.60	0.60
XGBoost	0.80	0.72	0.88	0.81
Logistic regression	0.60	0.64	0.75	0.50
Random Forest	0.70	0.60	0.99	0.50

Рисунок 3.37 – Точності використаних моделей

В середньому найвищу точність має модель градієнтного бустингу xgboost. Найточнішою в одиночному випадку є модель Random Forest з точністю прогнозу 99%, використана на датасеті з 6 семестру навчання [1].

3.2 Висновки

У третьому розділі побудовано моделі з використанням методу дерева рішень для виявлення закономірностей факторів, що впливають на рівень знань іноземної мови студентами закладів вищої освіти.

Запропоновані моделі дерев рішень виявили, що:

– «Латинська мова та медична термінологія» є найбільш впливовою на рівень вивчення іноземної мови студентами 2 семестру навчання;

– «Медична інформатика» є найбільш впливовою на рівень вивчення іноземної мови студентами 4 семестру навчання;

– «Загальна хірургія» є найбільш впливовою на рівень вивчення іноземної мови студентами 6 семестру навчання.

Також визначено, що в середньому найвищу точність має модель градієнтного бустингу xgboost. Але найточнішою в одиночному випадку є модель Random Forest з точністю прогнозу 99%, використана на датасеті з даними 6 семестру навчання

4 РЕАЛІЗАЦІЯ ПРОГРАМНОГО МОДУЛЯ АНАЛІЗУ РІВНЯ ЗНАНЬ ІНОЗЕМНОЇ МОВИ СТУДЕНТІВ ЗАКЛАДУ ВИЩОЇ ОСВІТИ

4.1 Реалізація клієнтської частини

Також у роботі було розроблено веб-систему для подальшого визначення рівня знань іноземної мови студентів закладу вищої освіти.

Перед початком розробки підключимо бібліотеки Bootstrap. На рисунку 4.1 зображений фрагмент коду, що використовується для підключення.

```
    <!-- Latest compiled and minified CSS -->
<link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.7/css/bootstrap.min.css"
integrity="sha384-BVYiiSIFeK1dGmJRAkycuHAHRg32OmUcww7on3RYdg4Va+PmSTsz/K68vbdEjh4u" crossorigin="anonymous">
<link rel="stylesheet" href="https://use.fontawesome.com/releases/v5.8.1/css/all.css"
integrity="sha384-50oBUHEmvpQ+1lW4y57PTFmhCaXp0ML5d60M1M7uH2+nqUivzIebhndOJK28anvf" crossorigin="anonymous">
```

Рисунок 4.1 – Підлючення бібліотеки Bootstrap

Отож, першим кроком у реалізації стане створення веб-сторінок, які в майбутньому будуть виводити інформацію з бази даних.

Почнемо з створення, так званої «шапки» сайту. У ній буде розташована навігація по сайту. Використаємо класи bootstrap для створення меню та випадаючого списку. Зовнішній вигляд створеної «шапки» зображено на рисунку 4.2.

Рисунок 4.2 – «Шапка» сайту

Код її створення продемонстровано на рисунку 4.3.

```
<nav class="navbar navbar-default navbar-fixed-top">
  <div class="container-fluid">
    <div class="navbar-header">
      <a class="navbar-brand" href="#">Englando</a>
    </div>
    <ul class="nav navbar-nav">
      <li ><a href="#">Home</a></li>
    </ul>
    <ul class="nav navbar-nav navbar-right">
      <li><a href="#" class="dropdown-toggle" data-toggle="dropdown"><span class="glyphicon glyphicon-user"></span> User</a>
        <ul class="dropdown-menu">
          <li><a href="#">Progress page</a></li>
          <li><a href="#">Settings</a></li>
          <li><a href="#">Exit</a></li>
      </li>
    </ul>
  </div>
</nav>
```

Рисунок 4.3 – Код для створення «шапки» сайту

Наступним кроком розробки системи, буде створення сторінки для авторизації користувачів. Вона складається з двох полів введення та кнопки для перевірки (рис. 4.4).

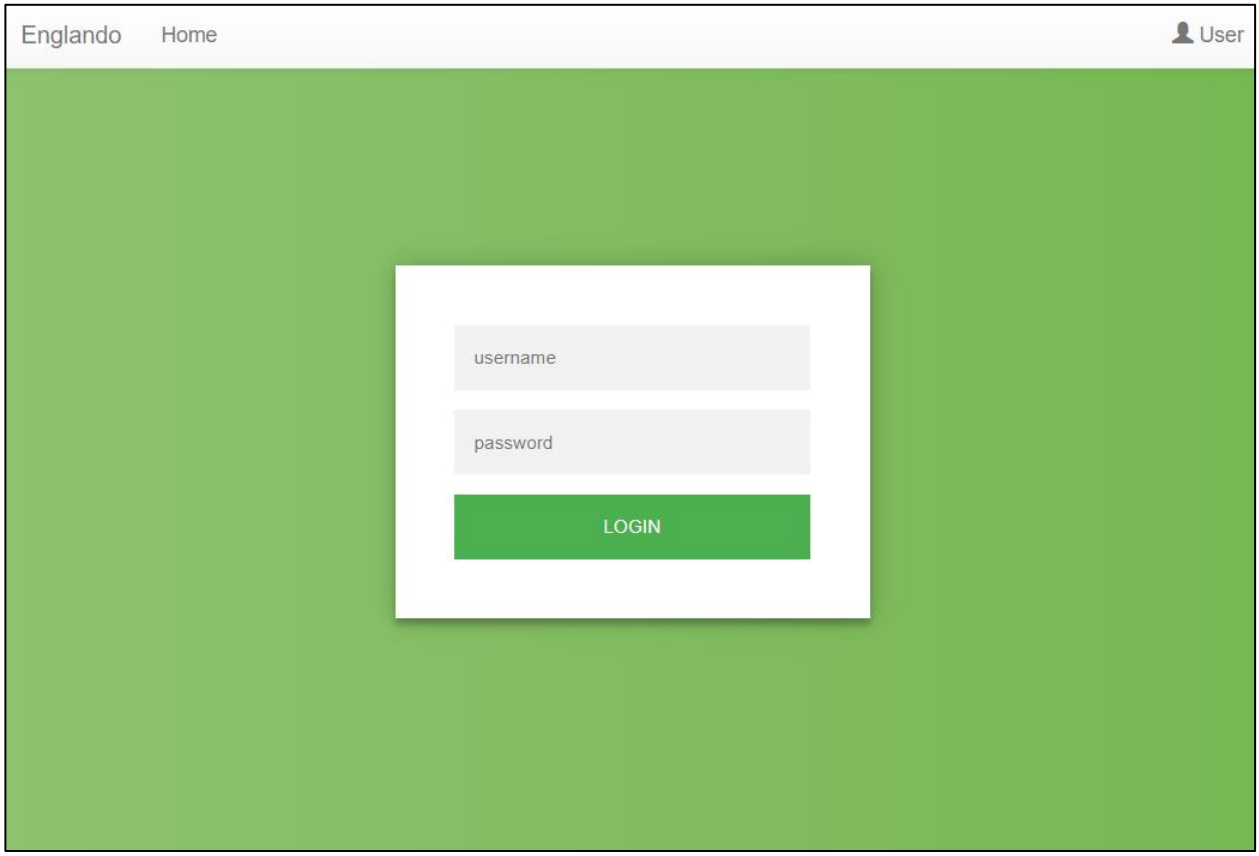

Рисунок 4.4 – Сторінка авторизації

HTML та CSS код цієї форми зображено на рисунках 4.5, 4.6 відповідно.

```
  <div class="login-page" style= "margin-top: 200px;">
<div class="form">
  <form class="login-form" method="POST">
    <input type="text" name="login" placeholder="username"/>
    <input type="password" name="password" placeholder="password"/>
    <button name="submit" type="submit">login</button>
  </form>
</div>
</div>
```

Рисунок 4.5 – Код для створення форми авторизації

```
.login-page {
  width: 360px;
  padding: 8% 0 0;
  margin: auto;
}
.form {
  position: relative;
  z-index: 1;
  background: #FFFFFF;
  max-width: 360px;
  margin: 0 auto 100px;
  padding: 45px;
  text-align: center;
  box-shadow: 0 0 20px 0 rgba(0, 0, 0, 0.2), 0 5px 5px 0 rgba(0, 0, 0, 0.24);
}
.form input {
  font-family: "Roboto", sans-serif;
  outline: 0;
  background: #f2f2f2;
  width: 100%;
  border: 0;
  margin: 0 0 15px;
  padding: 15px;
  box-sizing: border-box;
  font-size: 14px;
}
.form button {
  font-family: "Roboto", sans-serif;
  text-transform: uppercase;
  outline: 0;
  background: #4CAF50;
  width: 100%;
  border: 0;
  padding: 15px;
  color: #FFFFFF;
  font-size: 14px;
  -webkit-transition: all 0.3 ease;
  transition: all 0.3 ease;
  cursor: pointer;
}
.form button:hover,.form button:active,.form button:focus {
  background: #43A047;
```

Рисунок 4.6 – Код для створення стилю форми авторизації

Після авторизації користувач обирає тип завданнь (рис. 4.7).

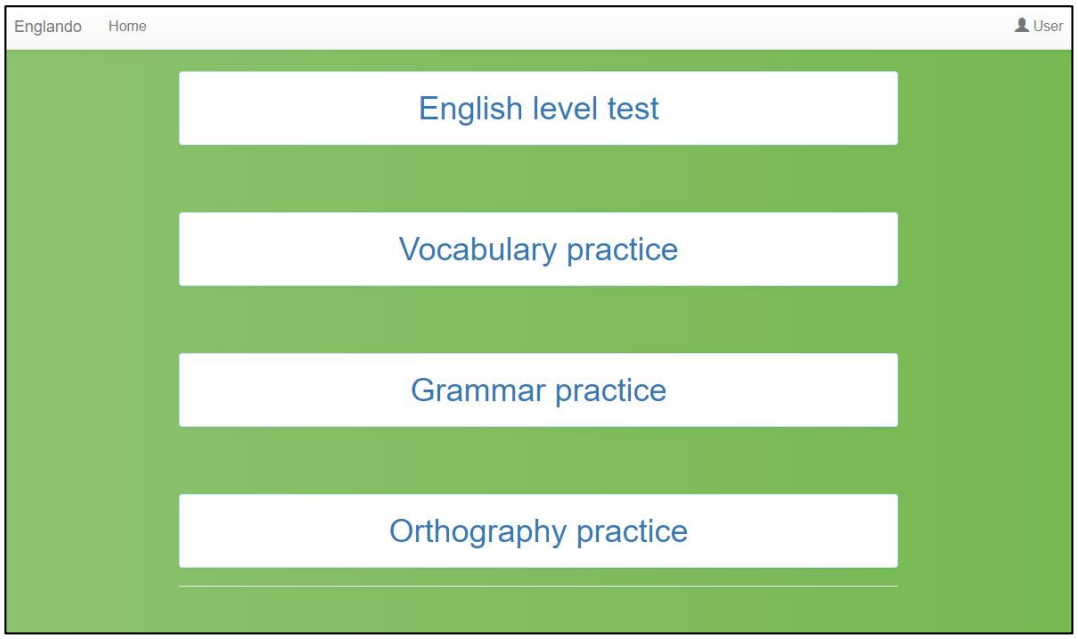

Рисунок 4.7 – Вибір типу завдань

Основним завданням являється «English level test» (рис 4.8), що являє собою тест на визначення рівня знань іноземної мови.

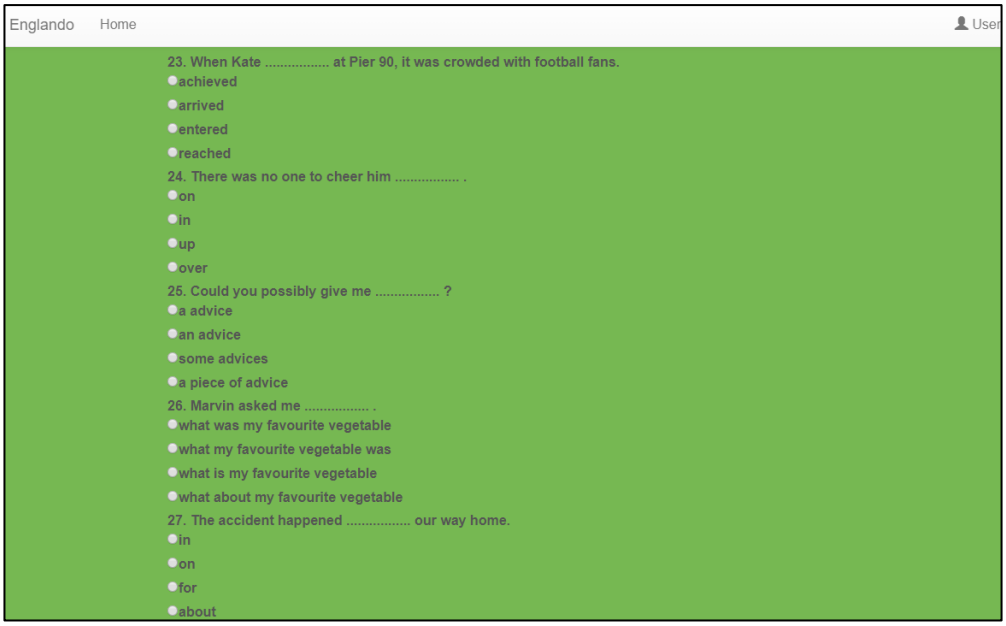

Рисунок 4.8 – Фрагмент «English level test»

Всі інші типи завдань поділяються на блоки з різних тем, що дозволяють підняти свій рівень знань іноземної мови.

На рисунку 4.9 зображено блоки типу «Vocabulary practice».

Рисунок 4.9 – Вибір блоку завдань

Приклад завдання блоку «Idioms» типу «Grammar practice» (рис. 4.10)

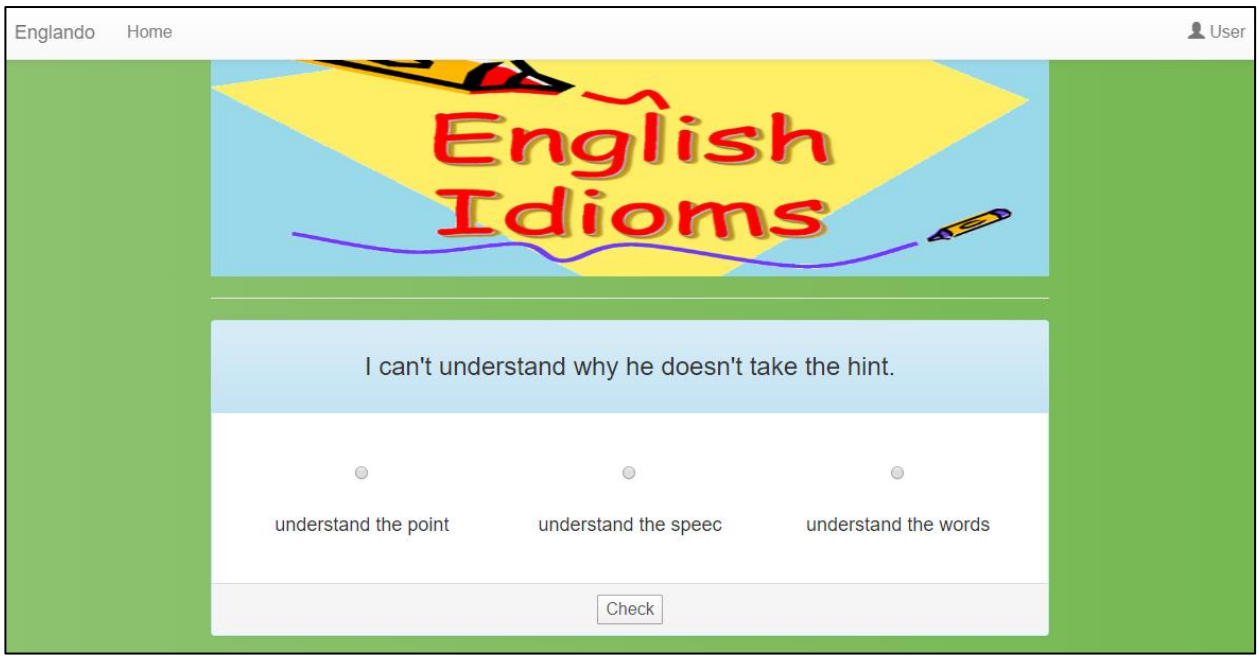

Рисунок 4.10 – Приклад завдання блоку «Idioms» типу «Grammar practice»

4.2 Реалізація серверної частини

Таким чином, було створено усі сторінки веб-системи. Але для того, щоб підключити базу даних, потрібно виконати ще кілька кроків. Структуру бази даних зображено на рисунку 4.11.

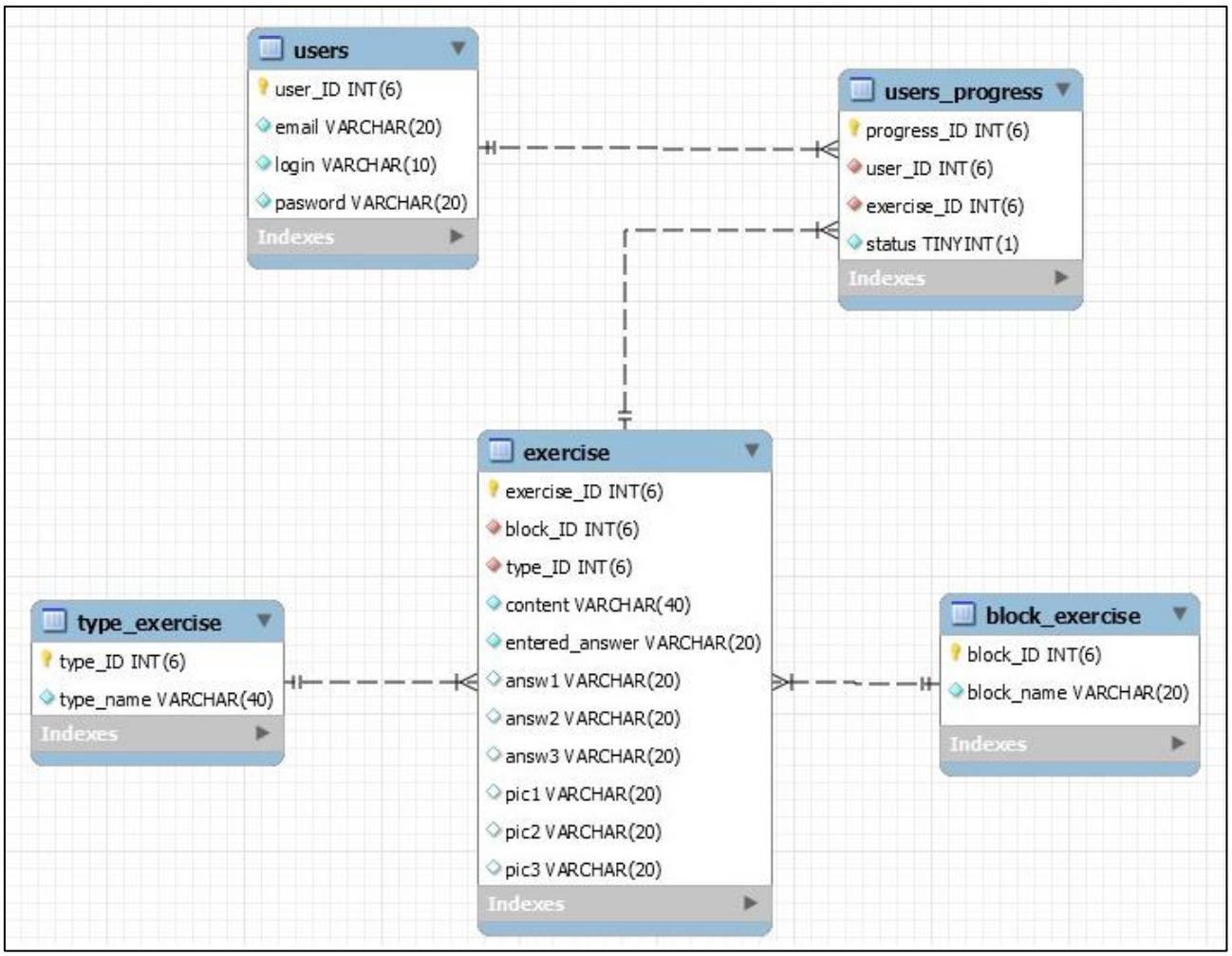

Рисунок 4.11 – Схема бази даних

База даних складається з 5 зв`язаних таблиць. Таблиці block_exercise та type_exercise, за своєю будовою, являються довідковими. В них зберігається тема завдання, або, так званий блок, та тип завдання, відповідно. Таблиця users містить дані користувача. Exercise являє собою, таблицю в якій зберігається

текст завдання, варіанти відповіді та правильна відповідь. Прогрес користувачів по кожному завданню міститься у таблиці users_progress.

На рисунку 4.12 продемонстровано SQL код для створення таблиці exercise.

```
CREATE TABLE IF NOT EXISTS `englando`.`exercise` (
  `exercise_ID` INT(6) NOT NULL AUTO_INCREMENT,
  `block_ID` INT(6) NOT NULL,
  `type_ID` INT(6) NOT NULL,
  `content` VARCHAR(40) NOT NULL,
  `entered_answer` VARCHAR(20) NOT NULL,
  `answ1` VARCHAR(20) NULL DEFAULT NULL,
  `answ2` VARCHAR(20) NULL DEFAULT NULL,
  `answ3` VARCHAR(20) NULL DEFAULT NULL,
  `pic1` VARCHAR(20) NULL DEFAULT NULL,
  `pic2` VARCHAR(20) NULL DEFAULT NULL,
  `pic3` VARCHAR(20) NULL DEFAULT NULL,
  PRIMARY KEY (`exercise_ID`),
  INDEX `block_ID` (`block_ID` ASC) VISIBLE,
  INDEX `type_ID` (`type_ID` ASC) VISIBLE,
  CONSTRAINT `exercise_ibfk_1`
    FOREIGN KEY (`block_ID`)
    REFERENCES `englando`.`block_exercise` (`block_ID`)
    ON DELETE CASCADE
    ON UPDATE CASCADE,
  CONSTRAINT `exercise_ibfk_2`
    FOREIGN KEY (`type_ID`)
    REFERENCES `englando`.`type_exercise` (`type_ID`)
    ON DELETE CASCADE
    ON UPDATE CASCADE)
ENGINE = InnoDB
```

Рисунок 4.12 – SQL код для створення таблиці exercise

Таким чином було створено усі таблиці бази даних. Але для того, щоб підключити базу даних, потрібно виконати ще кілька кроків.

Встановимо локальний сервер Denwer. На рисунку 4.13 зображено запущений Denwer.

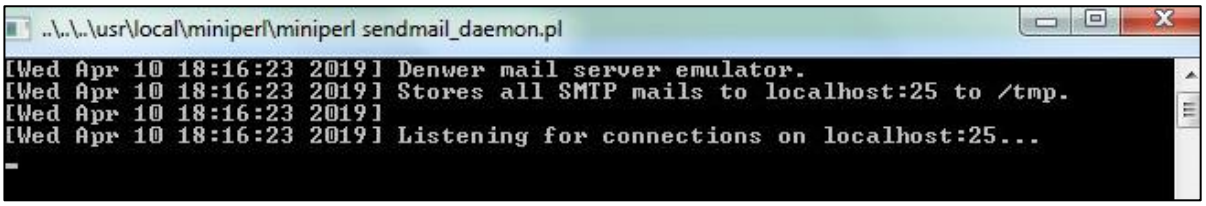

Рисунок 4.13 – Запущений Denwer

Для того, щоб відкрити phpMyAdmin перейдемо по посиланню localhost/tools/phpmyadmin/ (рис 4.14) [20].

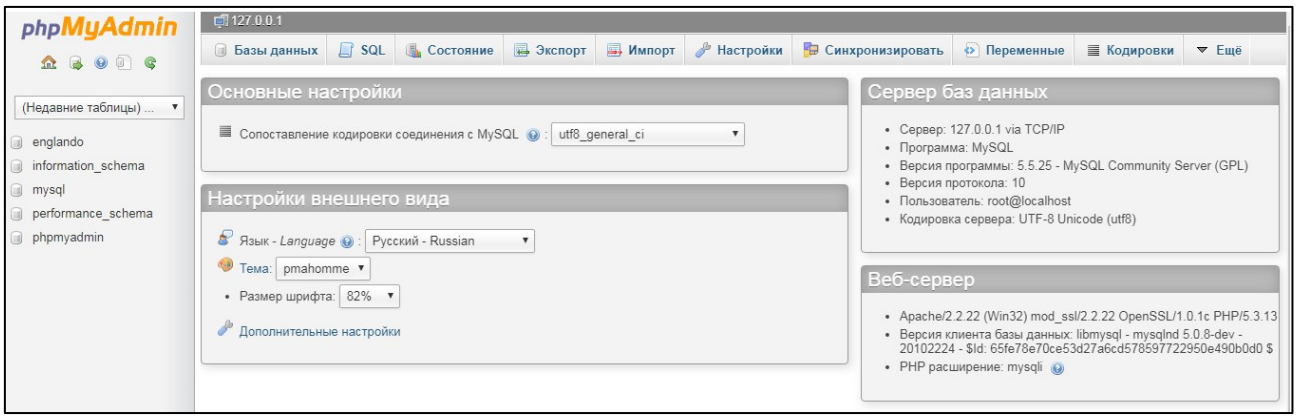

Рисунок 4.14 – Головна сторінка phpMyAdmin

Тепер перейдемо на вкладку SQL та вставимо код спроектованої бази даних (рис. 4.15).

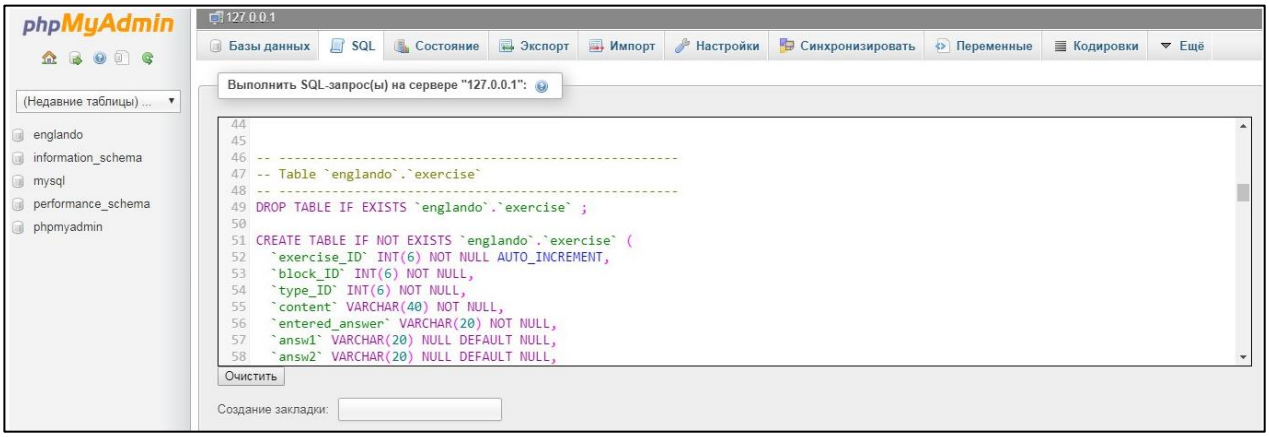

Рисунок 4.15 – Створення бази даних в phpMyAdmin

Заповнимо базу даними для подальшого підключення до сайту (рис. 4.16).

←T→	exercise_ID	block_ID	type_ID	content	entered_answer	answ1	answ2	answ3	pic1	pic2	pic3
Изменить Копировать Удалить	1	1	1	Translate "apple"	яблуко	груша	яблуко	огірок	NULL	NULL	NULL
Изменить Копировать Удалить	2	1	1	Виберіть переклад слова "cucumber"	огірок	груша	яблуко	огірок	NULL	NULL	NULL
Изменить Копировать Удалить	6	2	1	Виберіть переклад слова "gloves"	рукавиці	капелюх	футболка	рукавиці	NULL	NULL	NULL
Изменить Копировать Удалить	4	2	1	Виберіть переклад слова "hat"	капелюх	капелюх	футболка	рукавиці	NULL	NULL	NULL
Изменить Копировать Удалить	3	1	1	Виберіть переклад слова "pear"	груша	груша	яблуко	огірок	NULL	NULL	NULL
Изменить Копировать Удалить	5	2	1	Виберіть переклад слова "t-shirt"	футболка	капелюх	футболка	рукавиці	NULL	NULL	NULL

Рисунок 4.16 – Таблиця exercise з даними

Підключення до бази даних виконано за допомогою php. На рисунку 4.17 продемонстровано фрагмент коду, який формує зв`язок між базою та веб-сторінкою.

```
<?php
$host = "localhost";
$database = "englando";
$user = "root";
$password = "";
```

Рисунок 4.17 – Підключення до бази даних

Тепер можемо вивести потрібну нам інформацію з таблиць бази даних, на веб-сторінки. Інформація обирається запитами SQL (рис. 4.18)

```
$mysqli = new mysqli ("localhost", "root", "", "englando");
$query = "set names utf8";
$mysqli->query($query);
$query = "select * from block_exercise ";
$results = $mysqli->query($query);
```

Рисунок 4.18 – Вибірка даних з БД

Приклад коду, що виводить дані з БД зображено на рисунку 4.19.

```
while($row = $results->fetch_assoc()){

            echo '
    <div class="panel-group 3">
    <div class="panel panel-info">
    <div class="panel-body centred"><a href="index.php?id='.$row["block_ID"].'">'.$row["block_name"].'</a></div>

    </div>
    </div>
';


  }
```

Рисунок 4.19 – Код для виведення даних на HTML-сторінку

Для перевірки правильності вибору відповіді виконується перевірка (рис. 4.20).

```
$(document).ready(function(){
$("input[type='button']").click(function(){
    var id = this.id;
    num_quest++;
    var radioValue = $("input[name='" + id + "']:checked").val();
    if(radioValue === answ[id]){
        alert("Your are right, it is - " + radioValue);
        num_answ++;
    }else{alert("Your are wrong");}
    this.disabled = true;
        });
    });
```

Рисунок 4.20 – Код для перевірки вибору

Перевірка проводиться після натиснення на кнопку. Після проходження тесту виводиться статистика відповідей на питання (рис. 4.21).

Рисунок 4.21 – Результат проходження тесту

4.3 Висновки

Результатом виконання цього розділу є програмний модуль у вигляді веб-системи з використанням основних веб-технологій: мови гіпертекстової розмітки HTML, каскадні таблиці CSS та шаблони бібліотеки Bootstrap для побудови макету сайту, локальний сервер Denwer, веб-інтерфейс phpMyAdmin, для роботи з базою даних, написоною мовою запитів SQL. Також, було використано PHP для підключення макету сайту з базою даних. Програмний модуль включає в себе веб-сайт з підключеною базою даних.

5 ЕКОНОМІЧНА ЧАСТИНА

5.1 Оцінювання комерційного потенціалу розробки

Метою проведення комерційного та технологічного аудиту є оцінювання комерційного потенціалу впровадження алгоритму розвідувального аналізу та прогнозування результатів оцінювання студентів.

Для проведення технологічного аудиту було залучено 3-х незалежних експертів Вінницького національного технічного університету кафедри системного аналізу та інформаційних технологій: к.т.н., доц. Жуков С.О., к.т.н., доц. Козачко О.М., к.т.н., доц. Варчук І.В. Для проведення технологічного аудиту було використано таблицю 5.1 в якій за п’ятибальною шкалою використовуючи 12 критеріїв здійснено оцінку комерційного потенціалу [21].

Таблиця 5.1 – Рекомендовані критерії оцінювання комерційного потенціалу розробки та їх можлива бальна оцінка

Критерії оцінювання та бали (за 5-ти бальною шкалою)					
Кри-терій	0	1	2	3	4
Технічна здійсненність концепції:					
1	Достовірність концепції не підтверджена	Концепція підтверджена експертними висновками	Концепція підтверджена розрахунками	Концепція перевірена на практиці	Перевірено роботоздатність продукту в реальних умовах
Ринкові переваги (недоліки):					
2	Багато аналогів на малому ринку	Мало аналогів на малому ринку	Кілька аналогів на великому ринку	Один аналог на великому ринку	Продукт не має аналогів на великому ринку
3	Ціна продукту значно вища за ціни аналогів	Ціна продукту дещо вища за ціни аналогів	Ціна продукту приблизно дорівнює цінам аналогів	Ціна продукту дещо нижче за ціни аналогів	Ціна продукту значно нижче за ціни аналогів
4	Технічні та споживчі властивості продукту значно гірші, ніж в аналогів	Технічні та споживчі властивості продукту трохи гірші, ніж в аналогів	Технічні та споживчі властивості продукту на рівні аналогів	Технічні та споживчі властивості продукту трохи кращі, ніж в аналогів	Технічні та споживчі властивості продукту значно кращі, ніж в аналогів

Продовження табл. 5.1

Критерії оцінювання та бали (за 5-ти бальною шкалою)					
Кри-терій	0	1	2	3	4
5	Експлуатаційні витрати значно вищі, ніж в аналогів	Експлуатаційні витрати дещо вищі, ніж в аналогів	Експлуатаційні витрати на рівні експлуатаційних витрат аналогів	Експлуатаційні витрати трохи нижчі, ніж в аналогів	Експлуатаційні витрати значно нижчі, ніж в аналогів
Ринкові перспективи					
6	Ринок малий і не має позитивної динаміки	Ринок малий, але має позитивну динаміку	Середній ринок з позитивною динамікою	Великий стабільний ринок	Великий ринок з позитивною динамікою
7	Активна конкуренція великих компаній на ринку	Активна конкуренція	Помірна конкуренція	Незначна конкуренція	Конкурентів немає
Практична здійсненність					
8	Відсутні фахівці як з технічної, так і з комерційної реалізації ідеї	Необхідно наймати фахівців або витрачати значні кошти та час на навчання наявних фахівців	Необхідне незначне навчання фахівців та збільшення їх штату	Необхідне незначне навчання фахівців	Є фахівці з питань як з технічної, так і з комерційної реалізації ідеї
9	Потрібні значні фінансові ресурси, які відсутні.	Потрібні незначні фінансові ресурси.	Потрібні значні фінансові ресурси.	Потрібні незначні фінансові ресурси.	Не потребує додаткового фінансування
10	Необхідна розробка нових матеріалів	Потрібні матеріали, що використовуються у військово-промисловому комплексі	Потрібні дорогі матеріали	Потрібні досяжні та дешеві матеріали	Всі матеріали для реалізації ідеї відомі та давно використовуються у виробництві

Продовження табл. 5.1

Критерії оцінювання та бали (за 5-ти бальною шкалою)					
Кри-терій	0	1	2	3	4
11	Термін реалізації ідеї більший за 10 років	Термін реалізації ідеї більший за 5 років. Термін окупності інвестицій більше 10-ти років	Термін реалізації ідеї від 3-х до 5-ти років. Термін окупності інвестицій більше 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій від 3-х до 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій менше 3-х років
12	Необхідна розробка регламентних документів та отримання великої кількості дозвільних документів на виробництво та реалізацію продукту	Необхідно отримання великої кількості дозвільних документів на виробництво та реалізацію продукту, що вимагає значних коштів та часу	Процедура отримання дозвільних документів для виробництва та реалізації продукту вимагає незначних коштів та часу	Необхідно тільки пові-домлення відповідним органам про виробництво та реалізацію продукту	Відсутні будь-які регламентні обмеження на виробництво та реалізацію продукту

Таблиця 5.2 – Рівні комерційного потенціалу розробки

Середньоарифметична сума балів СБ, розрахована на основі висновків експертів	Рівень комерційного потенціалу розробки
0-10	Низький
11-20	Нижче середнього
21-30	Середній
31-40	Вище середнього
41-48	Високий

В таблиці 5.3 наведено результати оцінювання експертами комерційного потенціалу розробки.

Таблиця 5.3 – Результати оцінювання комерційного потенціалу розробки

Критерії	Прізвище, ініціали, посада експерта		
	1. Жуков С.О.	2. Козачко О.М.	3. Варчук І.В.
	Бали, виставлені експертами:		
1	4	4	3
2	1	1	1
3	3	4	4
4	2	3	3
5	3	3	3
6	2	3	3
7	2	2	2
8	3	1	1
9	4	4	3
10	4	4	4
11	4	3	3
12	4	4	3
Сума балів	$СБ_1$=36	$СБ_2$=36	$СБ_3$=31
Середньоарифметична сума балів $\overline{СБ}$	$\overline{СБ} = \frac{\sum_1^3 СБ_i}{3} = \frac{36+36+31}{3} = 34$		

Середньоарифметична сума балів, розрахована на основі висновків експертів склала 34 бали, що згідно таблиці 5.2 вважається, що рівень комерційного потенціалу проведених досліджень є вище середнього.

Реалізація засобами машинного навчання на Python. Може бути використана у аналізі та розумінні зв'язків між вивченими дисциплінами, у результаті підвищити рівень та якість освіти.

5.2 Прогнозування витрат на виконання науково-дослідної роботи

Витрати, пов’язані з проведенням науково-дослідної роботи групуються за такими статтями: витрати на оплату праці, витрати на соціальні заходи, матеріали, паливо та енергія для науково-виробничих цілей, витрати на службові відрядження, програмне забезпечення для наукових робіт, інші витрати, накладні витрати.

1. Основна заробітна плата кожного із дослідників $З_0$, якщо вони працюють в наукових установах бюджетної сфери визначається за формулою:

$$З_0 = \frac{М}{Т_р} * t \text{ [грн]} \quad (5.1)$$

де $М$ – місячний посадовий оклад конкретного розробника (інженера, дослідника, науковця тощо), грн.;

$Т_Р$ – число робочих днів в місяці; приблизно $Т_Р \approx 21...23$ дні;

t – число робочих днів роботи дослідника.

Для розробки алгоритму розвідувального аналізу та прогнозування результатів оцінювання студентів необхідно залучити програміста з посадовим окладом 10000 грн. Кількість робочих днів у місяці складає 22, а кількість робочих днів програміста складає 22. Зведемо сумарні розрахунки до таблиця 5.5.

Таблиця 5.4 – Заробітна плата дослідника в науковій установі бюджетної сфери

Найменування посади	Місячний посадовий оклад, грн.	Оплата за робочий день, грн.	Число днів роботи	Витрати на заробітну плату грн.
Керівник	15000	681,8	5	3409
Програміст	10000	454,5	22	10000
Всього				13409

2. Розрахунок додаткової заробітної плати робітників

Додаткова заробітна плата $З_д$ всіх розробників та робітників, які приймали устають в розробці нового технічного рішення розраховується як 10 - 12 % від основної заробітної плати робітників.

На даному підприємстві додаткова заробітна плата начисляється в розмірі 10% від основної заробітної плати.

$$З_д = (З_о + З_р) * \frac{Н_{дод}}{100\%} \quad (5.2)$$

$$З_д = 0{,}11 * 13409 = 1475 \text{ (грн)}$$

3. Нарахування на заробітну плату $Н_{ЗП}$ дослідників та робітників, які брали участь у виконанні даного етапу роботи, розраховуються за формулою (5.3):

$$Н_{ЗП} = (З_о + З_д) * \frac{\beta}{100} \text{ [грн]} \quad (5.3)$$

де $З_о$ – основна заробітна плата розробників, грн.;

$З_д$ – додаткова заробітна плата всіх розробників та робітників, грн.;

β – ставка єдиного внеску на загальнообов'язкове державне соціальне страхування, % .

Дана діяльність відноситься до бюджетної сфери, тому ставка єдиного внеску на загальнообов'язкове державне соціальне страхування буде складати 22%, тоді**:**

$$Н_{ЗП} = (13409 + 1475) * \frac{22}{100} = 3274{,}5 \text{ (грн)}$$

4. Витрати на комлектуючі К, що були використані під час виконання даного етапу роботи, розраховуються по кожному виду матеріалів за формулою:

$$K = \sum_1^n Н_і \cdot Ц_і \cdot К_і, \quad (5.5)$$

де $Н_i$ – кількість комплектуючих *i*-го виду, шт.;

$Ц_i$ – покупна ціна комплектуючих *i*-го найменування, грн.;

$К_і$ – коефіцієнт транспортних витрат (1,1…1,15).

Таблиця 5.5 – Комплектуючі, що використані на розробку

Найменування матеріалу	Ціна за одиницю, грн.	Витрачено	Вартість витраченого матеріалу, грн.
Папір	211	1	211
Ручка	12	2	24
CD-диск	12	1	12
Флешка	135	1	135
Всього			382
З врахуванням коефіцієнта транспортування			420,2

5. Амортизація обладнання, комп’ютерів та приміщень, які використовувались під час виконання даного етапу роботи

Дані відрахування розраховують по кожному виду обладнання, приміщенням тощо.

$$А = \frac{Ц \cdot Н_а}{100} \cdot \frac{Т}{12} \quad (5.5)$$

де Ц – балансова вартість даного виду обладнання (приміщень), грн.;

$Т_{кор}$ – час користування;

Т – термін використання обладнання (приміщень), цілі місяці.

Згідно пункту 137.3.3 Податкового кодексу, амортизація нараховується на основні засоби вартістю понад 2500 грн. В нашому випадку для написання магістерської роботи використовувався персональний комп’ютер вартістю 35000 грн.

$$А = \frac{35000 \cdot 1}{2 \cdot 12} = 1458{,}33$$

6. До статті «Паливо та енергія для науково-виробничих цілей» відносяться витрати на всі види палива й енергії, що безпосередньо використовуються з технологічною метою на проведення досліджень.

$$В_е = \sum_{i=1}^{n} \frac{W_{yt} \cdot t_i \cdot Ц_е \cdot К_{впі}}{\eta_і} \quad (5.6)$$

де W_{yt} – встановлена потужність обладнання на певному етапі розробки, кВт;

t_i – тривалість роботи обладнання на етапі дослідження, год;

$Ц_е$ – вартість 1 кВт-години електроенергії, грн;

$К_{впі}$ – коефіцієнт, що враховує використання потужності, $К_{впі} < 1$;

$\eta_і$ – коефіцієнт корисної дії обладнання, $\eta_і < 1$.

Для написання магістерської роботи використовується персональний комп'ютер для якого розрахуємо витрати на електроенергію.

$$В_е = \frac{0{,}3 \cdot 180 \cdot 4{,}1 \cdot 0{,}5}{0{,}8} = 138{,}38$$

Витрати на службові відрядження, витрати на роботи, які виконують сторонні підприємства, установи, організації та інші витрати в нашому досліджені не враховуються оскільки їх не було.

Накладні (загальновиробничі) витрати охоплюють: витрати на управління організацією, оплата службових відряджень, витрати на утримання, ремонт та експлуатацію основних засобів, витрати на опалення, освітлення, водопостачання, охорону праці тощо. Накладні (загальновиробничі) витрати можна прийняти як (100…150)% від суми основної заробітної плати розробників та робітників, які виконували дану МКНР, тобто:

$$В_{нзв} = (З_о + З_р) \cdot \frac{Н_{нзв}}{100\%}, \quad (5.7)$$

де $Н_{нзв}$ – норма нарахування за статтею «Інші витрати».

$$В_{нзв} = 13409 \cdot \frac{100}{100\%} = 13409 \text{ грн}$$

Сума всіх попередніх статей витрат дає витрати, які безпосередньо стосуються даного розділу МКНР

$$В = 13409+1475+3274,5+420,2+1458,33+138,33+13409=33854,6 \text{ грн}$$

Прогнозування загальних втрат ЗВ на виконання та впровадження результатів виконаної МКНР здійснюється за формулою:

$$ЗВ = \frac{В}{\eta} \quad (5.8)$$

де η – коефіцієнт, який характеризує стадію виконання даної НДР.

Оскільки, робота знаходиться на стадії науково-дослідних робіт, то коефіцієнт β = 0,4. Звідси:

$$ЗВ = \frac{33584,6}{0,5} = 67169,18 \text{ грн.}$$

5.3 Розрахунок економічної ефективності науково-технічної розробки

У даному підрозділі кількісно спрогнозуємо, яку вигоду, зиск можна отримати у майбутньому від впровадження результатів виконаної наукової роботи. Розрахуємо збільшення чистого прибутку підприємства $\Delta П_i$, для кожного із років, протягом яких очікується отримання позитивних результатів від впровадження розробки, за формулою

$$\Delta П_i = \sum_1^n (\Delta П_я \cdot N + П_я \Delta N)_i, \quad (5.9)$$

де $\Delta Ц_0$ – покращення основного оціночного показника від впровадження результатів розробки у даному році.

N – основний кількісний показник, який визначає діяльність підприємства у даному році до впровадження результатів наукової розробки;

ΔN – покращення основного кількісного показника діяльності підприємства від впровадження результатів розробки:

$Ц_о$ – основний оціночний показник, який визначає діяльність підприємства у даному році після впровадження результатів наукової розробки;

n – кількість років, протягом яких очікується отримання позитивних результатів від впровадження розробки:

л – коефіцієнт, який враховує сплату податку на додану вартість. Ставка податку на додану вартість дорівнює 20%, а коефіцієнт *л* = 0,8333.

р – коефіцієнт, який враховує рентабельність продукту. р = 0,25;

х – ставка податку на прибуток. У 2021 році – 18%.

Припустимо, що при впроваджені результатів наукової розробки покращується якість програмного продукту для формування індивідуальних тренувань. Припустимо, що ціна від зросте на 500 грн. Кількість одиниць реалізованої продукції також збільшиться: протягом першого року на 200 шт., протягом другого року – на 150 шт., протягом третього року на 1000 шт. Реалізація продукції до впровадження розробки складала 1 шт., а її ціна до складає 3000 грн. Розрахуємо прибуток, яке отримає підприємство протягом трьох років.

$$\Delta П_1 = [500 \cdot 1 + (3000 + 500) \cdot 200] \cdot 0{,}833 \cdot 0{,}25 \cdot \left(1 + \frac{18}{100}\right) = 119663{,}96 \text{ грн.}$$

$$\Delta П_2 = [500 \cdot 1 + (3000 + 500) \cdot (200 + 150)] \cdot 0{,}833 \cdot 0{,}25 \cdot \left(1 + \frac{18}{100}\right) = 209762{,}46 \text{ грн.}$$

$$\Delta П_3 = [500 \cdot 1 + (3000 + 500) \cdot (200 + 150 + 100)] \cdot 0{,}833 \cdot 0{,}25 \cdot \left(1 + \frac{18}{100}\right) = 269551{,}74 \text{ грн.}$$

5.4 Розрахунок ефективності вкладених інвестицій та періоду їх окупності

Розрахуємо основні показники, які визначають доцільність фінансування наукової розробки певним інвестором, є абсолютна і відносна ефективність вкладених інвестицій та термін їх окупності.

Розрахуємо величину початкових інвестицій PV, які потенційний інвестор має вкласти для впровадження і комерціалізації науково-технічної розробки.

$$PV = k_{\text{інв}} \cdot \text{ЗВ}, \tag{5.10}$$

$k_{\text{інв}}$ – коефіцієнт, що враховує витрати інвестора на впровадження науково-технічної розробки та її комерціалізацію. Це можуть бути витрати на підготовку приміщень, розробку технологій, навчання персоналу, маркетингові заходи тощо ($k_{\text{інв}} = 2…5$).

$$PV = 2 \cdot 67169{,}18 \ = 134338{,}36$$

Розрахуємо абсолютну ефективність вкладених інвестицій $E_{абс}$ згідно наступної формули:

$$\text{Е}_{\text{абс}} = (\text{ПП} - \text{PV}) \tag{5.11}$$

де ПП – приведена вартість всіх чистих прибутків, що їх отримає підприємство від реалізації результатів наукової розробки, грн.;

$$ПП = \sum_1^T \frac{\Delta П_i}{(1+\tau)^t} \quad (5.12)$$

де $\Delta П_i$ – збільшення чистого прибутку у кожному із років, протягом яких виявляються результати виконаної та впровадженої НДЦКР, грн.;

Т – період часу, протягом якою виявляються результати впровадженої НДДКР, роки;

τ – ставка дисконтування, за яку можна взяти щорічний прогнозований рівень інфляції в країні; для України цей показник знаходиться на рівні 0,2;

t – період часу (в роках).

$$ПП = \frac{119663,96}{(1+0,2)^1} + \frac{209762,46}{(1+0,2)^2} + \frac{269551,74}{(1+0,2)^3} = 402104,47 \text{ грн.}$$

$$Е_{абс} = (402104,47 - 134338,36\ ) = 267766,12 \text{грн.}$$

Оскільки $Е_{абс} > 0$, то вкладання коштів на виконання та впровадження результатів НДДКР може бути доцільним.

Розрахуємо відносну (щорічну) ефективність вкладених в наукову розробку інвестицій $E_в$. Для цього користуються формулою:

$$Е_в = \sqrt[T]{1 + \frac{Е_{абс}}{PV}} - 1, \quad (5.13)$$

Де $Е_{абс}$ – абсолютна ефективність вкладених інвестицій, грн;

PV – теперішня вартість інвестицій PV = ЗВ, грн;

Т – життєвий цикл наукової розробки, роки.

$$E_в = \sqrt[3]{1 + \frac{267766,12}{134338,36}} - 1 = 0,71 = 71\%$$

Визначимо мінімальну ставку дисконтування, яка у загальному вигляді визначається за формулою:

$$\tau = d + f \quad (5.14)$$

де d – середньозважена ставка за депозитними операціями в комерційних банках; в 2021 році в Україні d= (0,14…0,2);

f−показник, що характеризує ризикованість вкладень; зазвичай, величина f = (0,04...0,1).

$$\tau_{min} = 0,18 + 0,05 = 0,23$$

Так як $E_в > \tau_{min}$ то інвестор може бути зацікавлений у фінансуванні даної наукової розробки.

Розрахуємо термін окупності вкладених у реалізацію наукового проекту інвестицій за формулою:

$$T_{ок} = \frac{1}{E_в}, \quad (5.15)$$

$$T_{ок} = \frac{1}{0,71} = 1,4 \text{ роки}$$

Так як $T_{ок}$< 3...5-ти років, то фінансування даної наукової розробки в принципі є доцільним.

5.5 Висновки

Було проведено оцінку комерційного потенціалу алгоритму розвідувального аналізу та прогнозування результатів оцінювання студентів, який є на вище середньому рівні.

Прогнозування витрат на виконання науково-дослідної роботи по кожній з статей витрат складе 33584,6 грн. Загальна ж величина витрат на виконання та впровадження результатів даної НДР буде складати 67169,18 грн.

Вкладені інвестиції в даний проект окупляться через 1,4 роки при прогнозованому прибутку 402104,7 грн. за три роки.

ВИСНОВКИ

В ході магістерської кваліфікаційної роботи було розроблено інформаційну технологію аналізу рівня знань іноземної мови студентів закладу вищої освіти . Проведено розвідувальний аналіз даних оцінювання студентів Вінницького національного медичного університету ім. М. І. Пирогова за 2, 4, 6 семестри та ЄДКІ з іноземної мови, та відібрано оптимальний набір ознак, за якими слід будувати модель. Проведено аналіз існуючих методів аналізу для розв'язання поставленої задачі. Побудовано матриці кореляції та графіки розподілу. Для побудови моделей використано методи бустингу бібліотек lightgbm та xgboost з використання мови Python. Побудовано діаграму важливості ознак та дерево рішень. Розроблений програмний код, прогнозує оцінку з іноземної мови.

В першому розділі проаналізовано інформаційні технології аналізу рівня знань іноземної мови студентів закладів вищої освіти. Дослідження показали, що існуючі методи недостатньо приділяють увагу виявленню закономірностей щодо факторів, які впливають на підвищення рівня знань іноземної мови. Тому в наступних розділах пропонується розробити такий метод.

В другому розділі розроблено метод аналізу початких даних, за допомогою якого визначено фактори, що впливають на рівень вивчення іноземної мови студентами Вінницького національного медичного університету ім. М. І. Пирогова.

В третьому розділі побудовано моделі з використанням методу дерева рішень для виявлення закономірностей факторів, що впливають на рівень знань іноземної мови студентами закладів вищої освіти. Запропоновані моделі дерев рішень виявили, що:

– «Латинська мова та медична термінологія» є найбільш впливовою на рівень вивчення іноземної мови студентами 2 семестру навчання;

– «Медична інформатика» є найбільш впливовою на рівень вивчення іноземної мови студентами 4 семестру навчання;

– «Загальна хірургія» є найбільш впливовою на рівень вивчення іноземної мови студентами 6 семестру навчання.

– «Патофізіологія», «Фізіологія» та «Англійська мова» за 6 семестр навчання є найбільш впливовими на результат здачі ЄДКІ з іноземної мови.

Також визначено, що в середньому найвищу точність має модель градієнтного бустингу xgboost. Але найточнішою в одиночному випадку є модель Random Forest з точністю прогнозу 99%, використана на датасеті з даними 6 семестру навчання

В четвертому розділі розроблено програмний модуль у вигляді веб-системи з використанням основних веб-технологій: мови гіпертекстової розмітки HTML, каскадні таблиці CSS та шаблони бібліотеки Bootstrap для побудови макету сайту, локальний сервер Denwer, веб-інтерфейс phpMyAdmin, для роботи з базою даних, написоною мовою запитів SQL. Також, було використано PHP для підключення макету сайту з базою даних. Програмний модуль надає можливості до визначення та покращення рівня знань іноземної мови.

В п’ятому розділі було виконано економічну частину кваліфікаційної роботи під час розрахунків було доведено:

– прогнозування витрат на виконання науково-дослідної роботи по кожній з статей витрат складе 33584,6 грн. Загальна ж величина витрат на виконання та впровадження результатів даної НДР буде складати 67169,18 грн.

– вкладені інвестиції в даний проект окупляться через 1,4 роки при прогнозованому прибутку 402104,7 грн. за три роки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Лотоцький А. О. Козачко О. М. ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ РІВНЯ ЗНАНЬ ІНОЗЕМНОЇ МОВИ СТУДЕНТІВ ЗАКЛАДУ ВИЩОЇ ОСВІТИ. *Всеукраїнська науково-практична інтернет-конференція студентів аспірантів та молодих науковців «Молодь в науці: дослідження, проблеми, перспективи (МН-2022)»*. 2021-2022. URL: https://conferences.vntu.edu.ua/index.php/mn/mn2022/paper/view/14221

2. Predicting Pupil's Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods. [Електронний ресурс]. URL: https://link.springer.com/chapter/10.1007/978-3-030-16621-2_58

3. Mathematics Learning through Computational Thinking Activities: A Systematic Literature Review. [Електронний ресурс]. URL: http://jucs.org/jucs_24_7/mathematics_learning_through_computational/jucs_24_07_0815_0845_barcelos.pdf

4. TEACHING COMPUTATIONAL THINKING TO ENGLISH LEARNERS. [Електронний ресурс]. URL: https://par.nsf.gov/servlets/purl/10073683

5. An overview of Machine Learning Technologies and their use in E-learning. [Електронний ресурс]. URL: https://ieeexplore.ieee.org/abstract/document/9151758

6. Predicting academic performance using automatic learning techniques: a review of the scientific literature. [Електронний ресурс]. URL: https://ieeexplore.ieee.org/abstract/document/9254065

7. Мількевич К. С., Редич О. В. Використання штучних нейронних мереж в оцінці показників діяльності закладів вищої освіти. *Збірник статей здобувачів вищої освіти другого (магістерського) рівня Навчально-наукового інституту інформаційних технологій. Серія: Наукові роботи студентів УДФСУ*. Ірпінь: УДФСУ. 2020. Випуск 23. С. 49-59.

8. Марценюк В. П., Семенець А. В., Стаханська О. О. Нейромережеве прогнозування складання студентами-медиками ліцензійного інтегрованого іспиту" крок 1" на основі результатів поточної успішності та семестрового

комплексного тестового. *Медична інформатика та інженерія*. 2010. Випуск 2. С. 57-63.

9. Кюрчев С.В., Кувачов В. П. Прогнозування успішності навчання студентів – один із напрямів підвищення якості освіти. *Збірник науково-методичних праць «Удосконалення освітньо-виховного процесу в закладі вищої освіти»*. Мелітополь: ТДАТУ. 2020. Випуск 23. С. 57-64.

10. Some Studies in Machine Learning Using the Game of Checkers [Електронний ресурс]. URL: https://ieeexplore.ieee.org/document/5392560

11. The changing science of machine learning. [Електронний ресурс]. URL: https://link.springer.com/article/10.1007%2Fs10994-011-5242-y

12. Приклади алгоритму дерева рішень у видобутку даних. [Електронний ресурс]. URL: https://uk.myservername.com/decision-tree-algorithm-examples-data-mining

13. Zhi-Hua Z. Ensemble Methods: Foundations and Algorithms. 2012. – 23 с.

14. A. Shinde, A. Sahu, D. Apley, G. Rungerю Preimages for Variation Patterns from Kernel PCA and Bagging.. 2014. – 46 с.

15. A Kaggler's Guide to Model Stacking in Practice [Електронний ресурс] URL: http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/

16. Philip M. Long, Rocco A. Servedio. Random classification noise defeats all convex potential boosters. 2010. – 304 с.

17. Module pandas_profiling. [Електронний ресурс] URL: https://pandas-profiling.github.io/pandas-profiling/docs/

18. Наказ МОЗ України від 05.10.2016 №1043 [Електронний ресурс]. URL: https://www.testcentr.org.ua/docs/MoH/MoH-2016.10.05-1043.pdf

19. Мокін В. Б., Лосенко А. В., Дратований М. В. Інтелектуальна технологія аналізу та передбачення цін на вживані автомобілі. *Вісник Вінницького політехнічного інституту*. Вінниця: ВНТУ. 2019. Випуск 6. С. 62-72.

20. Bringing MySQL to the web. [Електронний ресурс]. URL: https://www.phpmyadmin.net/docs/

21. Методичні вказівки до виконання економічної частини магістерських кваліфікаційних робіт / Уклад. : В. О. Козловський, О. Й. Лесько, В. В. Кавецький. – Вінниця : ВНТУ, 2021. – 42 с.

Додаток А
(обов’язковий)

Міністерство освіти і науки України
Вінницький національний технічний університет
Факультет комп’ютерних систем і автоматики

ЗАТВЕРДЖУЮ
Завідувач кафедри САІТ
_____ д.т.н., проф. Мокін В. Б.
«___» _____ 2021 р.

ТЕХНІЧНЕ ЗАВДАННЯ
на магістерську кваліфікаційну роботу
«ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ РІВНЯ ЗНАНЬ ІНОЗЕМНОЇ МОВИ СТУДЕНТІВ ЗАКЛАДУ ВИЩОЇ ОСВІТИ»
08-53.МКР.006.02.000.ТЗ

Керівник: к.т.н., доцент
_____ Козачко О. М.
«___» _____ 2021 р.

Розробив: студент гр. 2ІСТ-20м
_____ Лотоцький А. О .
«___» _____ 2021 р.

Вінниця 2021

1. Підстава для проведення робіт

Підставою для виконання роботи є наказ № __ по ВНТУ від «__» _____ 2021 р., та індивідуальне завдання на МКР, затверджене протоколом № __ засідання кафедри САІТ від «__» _____ 2021 р.

2. Джерела розробки:

- Мількевич К. С., Редич О. В. Використання штучних нейронних мереж в оцінці показників діяльності закладів вищої освіти. *Збірник статей здобувачів вищої освіти другого (магістерського) рівня Навчально-наукового інституту інформаційних технологій. Серія: Наукові роботи студентів УДФСУ*. Ірпінь: УДФСУ. 2020. Випуск 23. С. 49-59.
- Мокін В. Б., Лосенко А. В., Дратований М. В. Інтелектуальна технологія аналізу та передбачення цін на вживані автомобілі. *Вісник Вінницького політехнічного інституту*. Вінниця: ВНТУ. 2019. Випуск 6. С. 62-72.

3. Мета і призначення роботи:

Розробка інформаційної технології аналізу рівня знань іноземної мови студентів закладу вищої освіти.

4.Вихідні дані для проведення робіт:

Датасети з даними про оцінювання студентів за 2, 4, 6 семестри навчання та результати проходження першого етапу ЄДКІ з англійської мови.

5.Методи дослідження:

– розвідувальний аналіз;

– моделювання системи.

6. Етапи роботи і терміни їх виконання:

1. Аналіз предметної області ..... ____ – ____
2. Розробка інформаційної технології..... ____ – ____
3. Реалізація інформаційної технології..... ____ – ____
4. Оформлення пояснювальної записки. ..... ____ – ____

7. Очікувані результати та порядок реалізації:

Отримання інформаційної технології аналізу рівня знань іноземної мови студентів закладу вищої освіти.

8. Вимоги до розробленої документації

Пояснювальна записка оформлена у відповідності до вимог «Методичних вказівок до виконання та оформлення магістерських кваліфікаційних робіт для студентів спеціальності 126 – «Інформаційні системи та технології» денної форми навчання».

9. Порядок приймання роботи

Публічний захист ..... «___» _____.2021 р.
Початок розробки..... «___» _____ 2021 р.
Граничні терміни виконання МКР ..... «___» _____ 2021 р.

Розробив студент групи 2ІСТ-20м _____ Лотоцький А.О.

Додаток Б
(обов'язковий)

Протокол перевірки кваліфікаційної роботи

Назва роботи: «Інформаційна технологія аналізу рівня знань іноземної мови студентів закладу вищої освіти»
Тип роботи: магістерська кваліфікаційна робота
Підрозділ: кафедра САІТ
Науковий керівник: Козачко О.М. к.т.н., доцент

Показники звіту подібності

Unicheck	
Оригінальність	92,9 %
Схожість	7,1 %

Аналіз звіту подібності (відмітити потрібне)

- ☐ Запозичення, виявлені у роботі, оформлені коректно і не містять ознак плагіату.
- ☐ Виявлені у роботі запозичення не мають ознак плагіату, але їх надмірна кількість викликає сумніви щодо цінності роботи і самостійності її автора. Роботу направити на доопрацювання.
- ☐ Виявлені у роботі запозичення є недобросовісними і мають ознаки плагіату та/або в ній містяться навмисні спотворення тексту, що вказують на спроби приховування недобросовісних запозичень.

Заявляю, що ознайомлений з повним звітом подібності, який був згенерований системою щодо роботи

Автор _____ Лотоцький А.О.
(підпис)

Опис прийнятого рішення
Робота допускається до захисту

Особа, відповідальна за перевірку _____ Жуков С. О.
(підпис)

Додаток В

(довідниковий)

Лістинг програмного коду.

```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import seaborn as sns
#import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
sns.set_style('darkgrid')
cutdatafull=pd.read_csv('cutdatafull.csv')
x = cutdatafull.iloc[:,:25]
y = cutdatafull.iloc[:,25]
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.20, random_state =
45)
import lightgbm as lgbm
import xgboost as xgb
Xtrain, Xval, Ztrain, Zval = train_test_split(x, y, test_size=0.2, random_state=0)
train_set = lgbm.Dataset(Xtrain, Ztrain, silent=False)
valid_set = lgbm.Dataset(Xval, Zval, silent=False)
params = {
        'boosting_type':'gbdt',
```

```
        'objective': 'regression',
        'num_leaves': 31,
        'learning_rate': 0.05,
        'max_depth': -1,
        'subsample': 0.8,
        'bagging_fraction' : 1,
        'max_bin' : 5000 ,
        'bagging_freq': 20,
        'colsample_bytree': 0.6,
        'metric': 'rmse',
        'min_split_gain': 0.5,
        'min_child_weight': 1,
        'min_child_samples': 10,
        'scale_pos_weight':1,
        'zero_as_missing': True,
        'seed':0,
    }

modelL = lgbm.train(params, train_set = train_set, num_boost_round=1000,
                    early_stopping_rounds=50, verbose_eval=10, valid_sets=valid_set)
fig =  plt.figure(figsize = (15,15))
axes = fig.add_subplot(111)
lgbm.plot_importance(modelL,ax = axes,height = 0.5)
plt.show();plt.close()
#%% split training set to validation set
data_tr  = xgb.DMatrix(Xtrain, label=Ztrain)
data_cv  = xgb.DMatrix(Xval   , label=Zval)
data_train = xgb.DMatrix(x)
data_test  = xgb.DMatrix(y)
evallist = [(data_tr, 'train'), (data_cv, 'valid')]
```

```
parms = {'max_depth':8, #maximum depth of a tree
        'objective':'reg:linear',
        'subsample':0.8,#SGD will use this percentage of data
        'lambda '  :4, #L2 regularization term,>1 more conservative
        'colsample_bytree ':0.9,
        'colsample_bylevel':1,
        'min_child_weight': 10}
modelx = xgb.train(parms, data_tr, num_boost_round=200, evals = evallist,
                early_stopping_rounds=30, maximize=False,
                verbose_eval=10)

print('score = %1.5f, n_boost_round
=%d.'%(modelx.best_score,modelx.best_iteration))
from sklearn.linear_model import  LogisticRegression

# creating a model
model = LogisticRegression()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the test set results
y_pred = model.predict(x_test)
from sklearn.metrics import confusion_matrix

# creating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# printing the confusion matrix
plt.rcParams['figure.figsize'] = (8, 8)
```

```
sns.heatmap(cm, annot = True, cmap = 'Greens')
plt.title('Confusion Matrix for Logistic Regression', fontweight = 30, fontsize = 20)
plt.show()
from sklearn.ensemble import RandomForestClassifier

# creating a model
model = RandomForestClassifier()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the x-test results
y_pred = model.predict(x_test)
from sklearn.metrics import confusion_matrix

# creating a confusion matrix
cm = confusion_matrix(y_test, y_pred)

# printing the confusion matrix
plt.rcParams['figure.figsize'] = (8, 8)
sns.heatmap(cm, annot = True, cmap = 'Reds')
plt.title('Confusion Matrix for Random Forest', fontweight = 30, fontsize = 20)
plt.show()
```

Додаток Г

(обов’язковий)

ІЛЮСТРАТИВНА ЧАСТИНА

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ РІВНЯ ЗНАНЬ ІНОЗЕМНОЇ МОВИ СТУДЕНТІВ ЗАКЛАДУ ВИЩОЇ ОСВІТИ

Виконав: студент гр. 2ІСТ-20м
_____ Лотоцький А. О.
«___» _____ 2021 р.

Керівник: к.т.н., доцент
_____ Козачко О. М.
«___» _____ 2021 р.

Нормоконтроль: к.т.н., доцент
_____ Жуков С. О.
«___» _____ 2021 р.

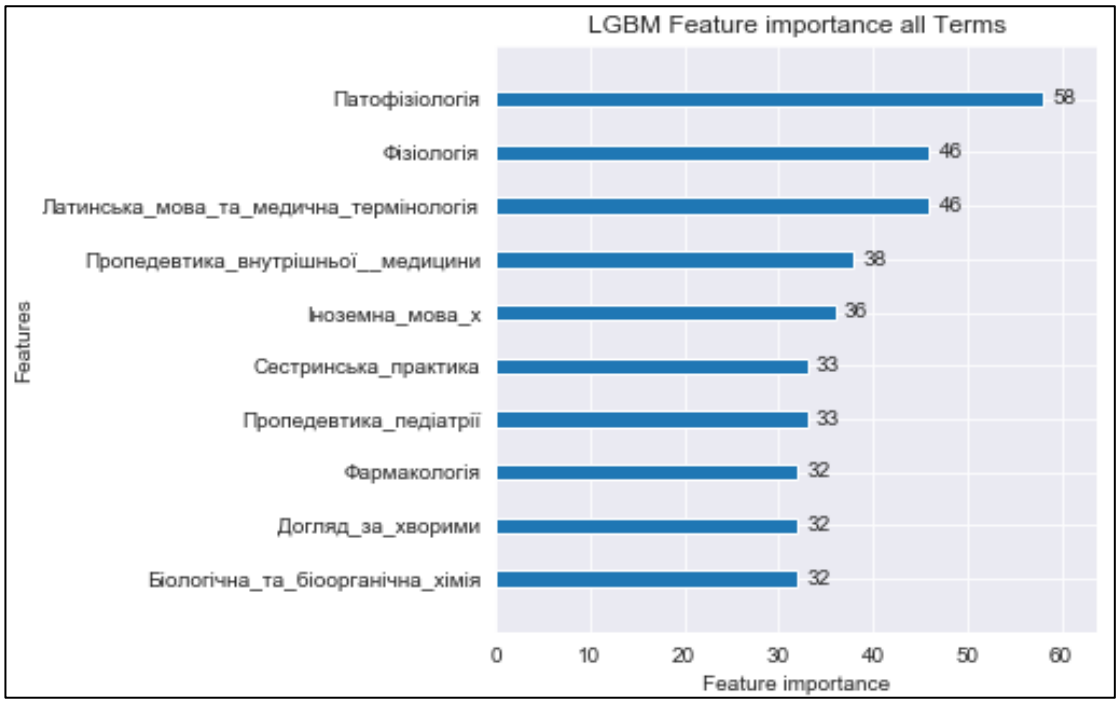

Рисунок Д.1 – Діаграма важливості ознак lightgbm моделі

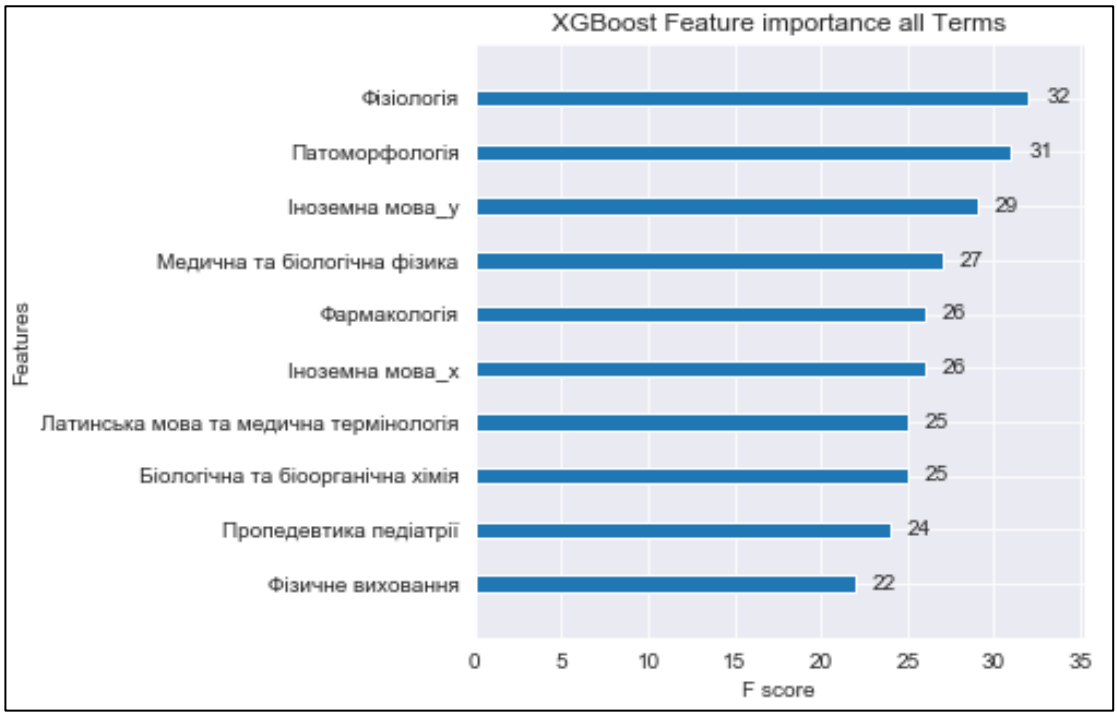

Рисунок Д.2 – Діаграма важливості ознак xgboost моделі

Рисунок Д.3 – Матриця невідповідності логістичної регресії

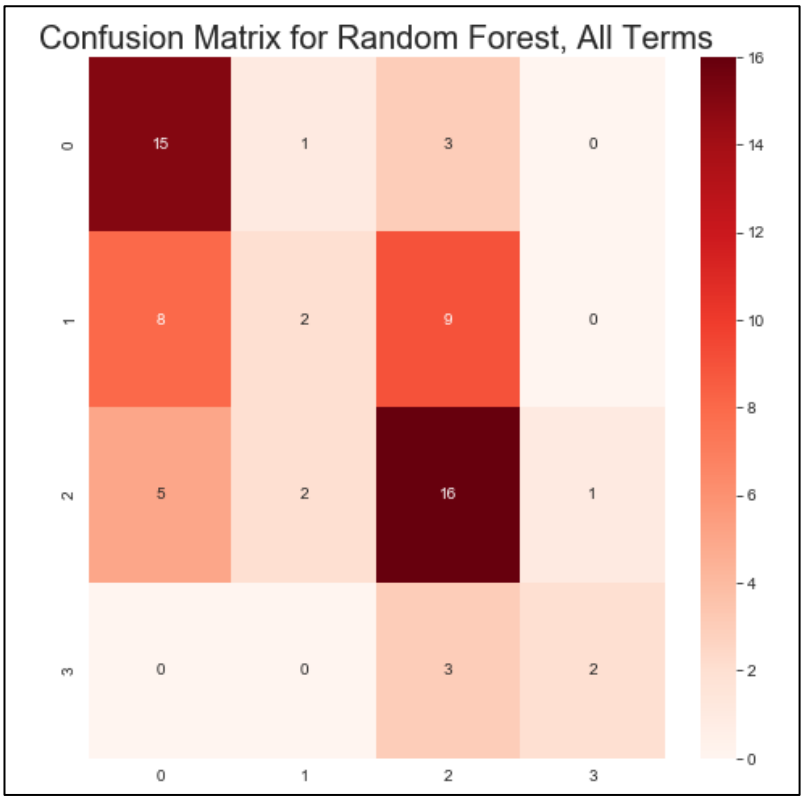

Рисунок Д.4 – Матриця невідповідності Random forest

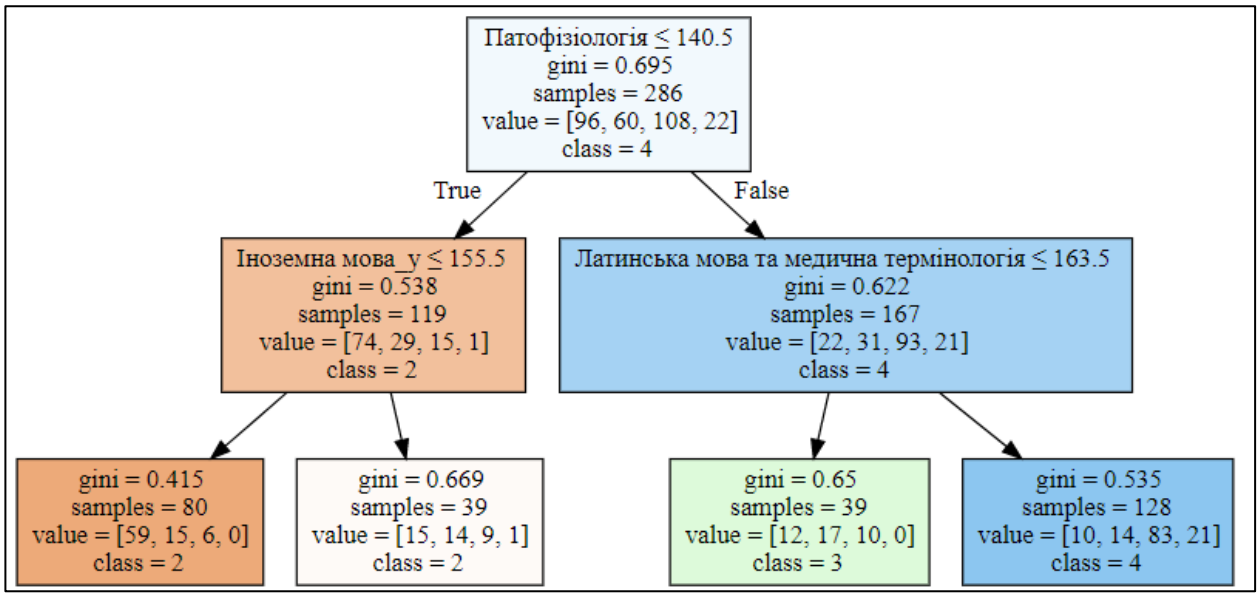

Рисунок Д.5 – Дерево рішень

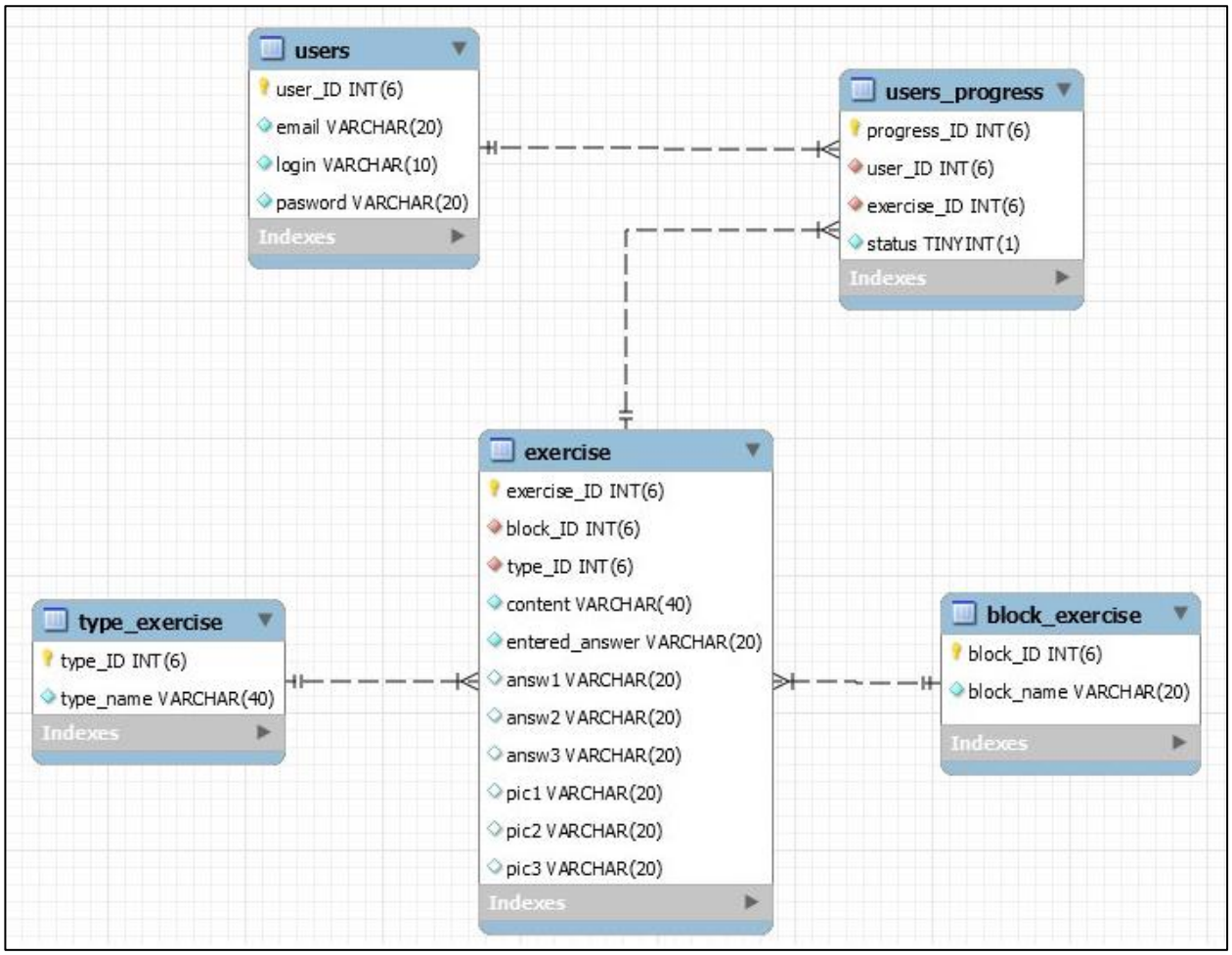

Рисунок Д.6 – Реляційна структура бази даних

Рисунок Д.7 – Загальний вигляд інтерфейсу веб-ресурсу