ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

Пояснювальна записка

до магістерської роботи
на ступінь вищої освіти магістр
на тему: «Розробка інформаційної технології для рекомендації туристичних маршрутів на основі методів машинного навчання»

Виконав: студент 6 курсу, групи ПДМ–61
спеціальності 121 Інженерія програмного забезпечення
(шифр і назва спеціальності/спеціалізації)

Трегубчак І.М.
(прізвище та ініціали)

Керівник Аверічев І.М.
(прізвище та ініціали)

Рецензент_____
(прізвище та ініціали)

Нормоконтроль_____
(прізвище та ініціали)

Київ –2022

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

Навчально-науковий інститут інформаційних технологій
Кафедра Інженерії програмного забезпечення
Ступінь вищої освіти -«Магістр»
Спеціальність підготовки – 121 «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Інженерії програмного забезпечення Негоденко О.В.
"____" _____2022року

З А В Д А Н Н Я
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

ТРЕГУБЧАКОВІ ІЛЛІ МИХАЙЛОВИЧУ

(прізвище, ім'я, по батькові)

1. Тема роботи: Розробка інформаційної технології для рекомендації туристичних маршрутів на основі методів машинного навчання

Керівник роботи: Аверічев Ігор Миколайович , к.е.н., доц., доцент кафедри ІПЗ ,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом вищого навчального закладу від «12» жовтня 2022 року №122.
2. Строк подання студентом роботи 31.12.2022

3. Вхідні дані до роботи
Історичні дані, які використовуються для навчання моделей машинного навчання.
Існуючі рекомендаційні системи
Науково-технічна література з питань, пов'язаних з розробкою рекомендаційних систем та побуди прогнозів за допомогою моделей машинного навчання.

4. Зміст розрахунково-пояснювальної записки(перелік питань, які потрібно розробити).
4.1 Системи аналізу, прогнозування та моделі машинного навчання
4.2 Вимоги та оцінка якості системи.
4.3 Опис математичних моделей та проектування системи.
4.4 Опис проектування системи.

5. Перелік демонстраційного матеріалу (назва основних слайдів)

1. Актуальність проблеми
2. Існуюче програмне забезпечення та методи прогнозування
3. Побудова прогнозу на основі обраних методів
4. Аналіз статистичних даних для проведення прогнозу
5. Моделювання процесу рекомендації маршруту
6. Проектування системи
7. Аналіз ефективності розробленої системи

6. Дата видачі завдання 14.10.2022

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Підбір науково-технічної літератури	14.10-25.10	
2	Вимоги до системи	28.10-05.11	
3	Оцінка якості тестування до системи	06.11-09.11	
4	Метод побудо	11.11-20.11	
5	Концепція та архітектура програмного забезпечення	21.11-30.11	
6	Вступ, висновки, реферат	30.11-05.11	
7	Розробка презентації	06.11-11.12	

Студент _____

Керівник роботи _____

РЕФЕРАТ

Текстова частина магістерської роботи 85 с., 20 рис., 24 джерел.

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ, ТУРИЗМ, МАШИННЕ НАВЧАННЯ, РЕКОМЕНДАЦІЙНІ СИСТЕМИ, ДЕРЕВО РІШЕНЬ, МЕТОД ОПОРНИХ ВЕКТОРІВ, БАГАТОШАРОВИЙ ПЕРСЕПТРОН.

Мета: покращення процесу вибору туристичних маршрутів за допомогою інформаційної технології на основі рекомендаційних систем з використанням методів машинного навчання.

Об’єкт: процес вибору туристичних маршрутів

Предмет: методи машинного навчання та рекомендаційні системи вибору туристичних маршрутів.

Методи дослідження: методи машинного навчання (дерева рішень, метод опорних векторів, багатошаровий персептрон), методи математичної статистики (мода, медіана, дисперсія), методи оптимізації.

Незважаючи на ситуацію в нашій країні туризм був і залишається однією з найпопулярніших сфер життя людини. Тому завжди залишається актуальним питання його оптимізації. Одним із таких шляхів оптимізації є застосування рекомендаційних систем.

Отже, розроблено та описано веб-додаток, завданням якого є на основі прогнозування та обробки даних за допомогою рекомендаційних систем оптимізувати процес вибору туристичного маршруту. У якості вихідних даних є історичні дані, які використовуються для навчання моделей машинного навчання..

Даний додаток може бути використано туристичними агенствами, а також звичайними користувачами, які планують відпочинок.

Галузь використання – туристична галузь.

ЗМІСТ

ВСТУП

Останнім часом туристична галузь модернізувалася за рахунок використання інформаційних технологій. Особливої популярності в туризмі набули інструменти підтримки прийняття рішень, також відомі як системи рекомендацій (СР). У сфері туризму вони називаються туристичними рекомендаційними системами (ТРС). За рахунок їх використання туристи та постачальники туристичних послуг можуть шукати, вибирати, порівнювати та приймати рішення майже миттєво та ефективніше, ніж будь-коли.

Через величезну кількість різнорідної інформації, доступної в Інтернеті та через інші джерела інформації, ТРС можуть діяти як інформаційні фільтри. Підбір відповідних туристичних послуг відповідно до вподобань користувачів є одним із найскладніших завдань, з якими стикається турист, плануючи поїздку в незнайоме місто. Незважаючи на те, що пошукові системи надають списки туристичних послуг, туристи все одно переповнені інформацією про існуючі пропозиції. ТРС можна легко використовувати як засіб зменшення інформаційного перевантаження для туристів.

ТРС можуть допомогти туристам самостійно подорожувати до незнайомого міста, особливо це стосується пошуку, вибору та порівняння туристичних послуг. Завдяки мобільному та бездротовому зв’язку ТРС можуть допомогти мандрівникам не лише під час планування подорожі, але й під час та після подорожі. Добре розвинена ТРС може запропонувати відповідні туристичні послуги туристам, не втручаючись у їхнє приватне життя, і запропонувати їм інші продукти, пов’язані з подорожами.

Крім того, ТРС можуть сприяти розвитку туризму в місті, а також рекламувати туристичні напрямки. Це матиме великий вплив на туризм міста чи країни, особливо туризм послуги, маркетинг і державні маркетингові стратегії. Що стосується компаній, пов’язаних із туризмом, то для того, щоб бути конкурентоспроможними та прибутковими та полегшити життя туристам, індустрія туризму та туристичні агенції повинні використовувати ТРС, щоб

гарантувати, що вони пропонують відмінні послуги туристам і таким чином покращують свій бізнес.

На сьогоднішній день більшість ТРС зосереджено на оцінках для вибору напрямків, заходів, пам'яток і туристичних послуг (наприклад, ресторанів, готелів і транспорту) на основі вподобань та інтересів користувачів. Що стосується технічних аспектів, ці ТРС забезпечують лише фільтрацію, сортування та основні механізми зіставлення між елементами та жорсткими та м'якими обмеженнями користувача. Щоб надавати туристам практичну допомогу, ТРС має стати «розумним» щодо певних технічних аспектів, таких як масштабованість, прозорість, точність рекомендацій і методи перевірки; а також певні практичні аспекти, такі як сприйняття користувачами та зручність використання – усі вони повинні бути взяті до уваги при проектуванні системи. Крім того, ефективна ТРС повинна досягти балансу між практичними та технічними аспектами.

У цьому дослідженні запропоновано розробити рекомендаційну систему, яка рекомендує туристам напрямки. Запропонована система прийняття рішень має дві основні відмінності порівняно з попередніми системами, які можна знайти в літературі: посилення системи за рахунок введення блоку машинного навчання та розробки зручного інтерфейсу користувача.

Мета: покращення процесу вибору туристичних маршрутів за допомогою інформаційної технології на основі рекомендаційних систем з використанням методів машинного навчання.

Об'єкт: процес вибору туристичних маршрутів

Предмет: методи машинного навчання та рекомендаційні системи вибору туристичних маршрутів.

Завдання:

1. Проаналізувати існуючі рекомендаційні системи, зокрема в сфері туризму. Визначити основні недоліки систем.
2. Визначення обмежень для системи.
3. Визначення вхідних даних та їх опрацювання.

4. Вибір методів машинного навчання та розробка алгоритму для їх ефективного застосування.

5. Розробка методики рекомендації туристичного маршруту.

6. Створення прототипу запропонованої інформаційної технології.

7. Визначення напрямків подальшого удосконалення проведеного дослідження.

Методи дослідження: методи машинного навчання (дерева рішень, метод опорних векторів, багатошаровий персептрон), методи математичної статистики (мода, медіана, дисперсія), методи оптимізації.

Наукова новизна: розроблена інформаційна технологія забезпечення пошуку оптимального туристичного напрямку на основі рекомендаційних систем, що включають в себе використання машинного навчання на нормалізації прогнозованих даних.

Практичне значення: побудовану інформаційну технологію можна використовувати як туристам для пошуку оптимального варіанту відпочинку, так і туроператорам для збільшення швидкості обробки запиту клієнтів.

Запропоновані результати дослідження були апробовані на XV Науково-технічній конференції «Сучасні інфокомунікаційні технології» в Державному університеті телекомунікацій.

За результатами отриманих досліджень були опубліковані наступні матеріали:

1. Трегубчак І.М. Методика вибору туристичних маршрутів з використанням методів машинного навчання // «ТІТ». №2, 2022

2. Трегубчак І.М. Методика вибору та обробки даних для побудови моделі рекомендаційної системи в сфері туризму // XV Науково-технічна конференція «Сучасні інфокомунікаційні технології» . – Київ: ДУТ, 2022.

1 СУЧАСНИЙ СТАН РЕКОМЕНДАЦІЙНИХ ТУРИСТИЧНИХ СИСТЕМ

1.1 Стан туризму в Україні та світі

Туризм – одна з небагатьох галузей, яка впродовж довгого періоду часу користувалася неймовірним попитом і тенденцією до зростання.

Поява СОВІД-19 ефективно вплинула на розвиток туристичної галузі, управління високотехнологічними інструментами управління та людськими ресурсами. В зв'язку з цим з'являються нові, технічні фактори, сучасні інноваційні механізми, наявні інтелектуальні ресурси, розвиток електронного бізнесу, розвиток багатьох туристичних видів – ділового, зеленого, сільськогосподарського, історико-культурного тощо. Хоча в наслідок пандемії показники туристичної діяльності значно скоротилися.

Проте за останніми розрахунками частка туризму в Україні становить близько 9% ВВП.

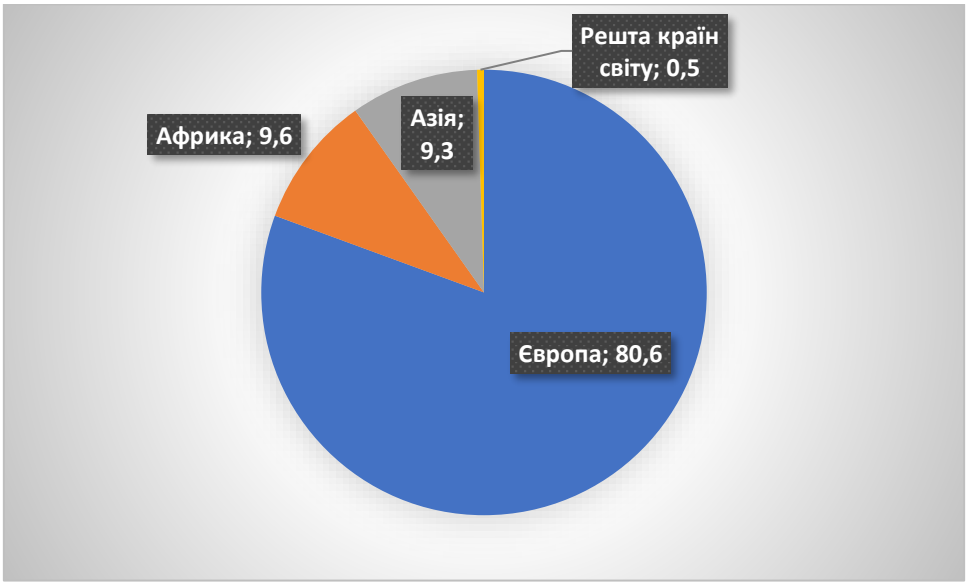


Рисунок 1.1 – Діаграма туристичних напрямів 2021

Таблиця 1.4 – Чисельність туристів у найпопулярніші туристичні країни 2021

№	Назва країни	Всього туристів	%Δ	Всього туристів
1.	Польща	3 301 511	-16.7%	30,28%
2.	Туреччина	1 737 848	80.0%	15,94%
3.	Угорщина	1 260 498	-22.9%	11,56%
4.	Єгипет	1 023 333	40.3%	9,38%
5.	Росія	649 215	-37.6%	5,95%
6.	Румунія	647 262	3.3%	5,94%
7.	Молдова	253 284	-23.0%	2,32%
8.	Словаччина	228 885	-32.0%	2,10%
9.	Німеччина	175 100	-21.2%	1,61%
10.	Білорусь	169 085	-65.9%	1,55%
11.	Греція	163 942	516.4%	1,50%
12.	ОАЕ	127 294	27.8%	1,17%
13.	Грузія	126 989	277.8%	1,16%
14.	Чорногорія	121 313	428.2%	1,11%
15.	Італія	105 863	77.1%	0,97%
16.	Болгарія	84 531	101.1%	0,78%
17.	Кіпр	62 780	326.5%	0,58%
18.	Австрія	56 764	-16.7%	0,52%
19.	Чехія	53 610	-1.3%	0,49%
20.	Нідерланди	50 129	20.7%	0,46%
21.	Іспанія	44 316	-17.3%	0,41%
22.	Домініканська Республіка	44 262	205.0%	0,41%
	Загальний висновок	**10 904 744**	**-3.1%**	**100.0%**

В таблиці 1.4 та на рис. 1.2 наведені числові дані по розподілу туристів на основних туристичних напрямках.

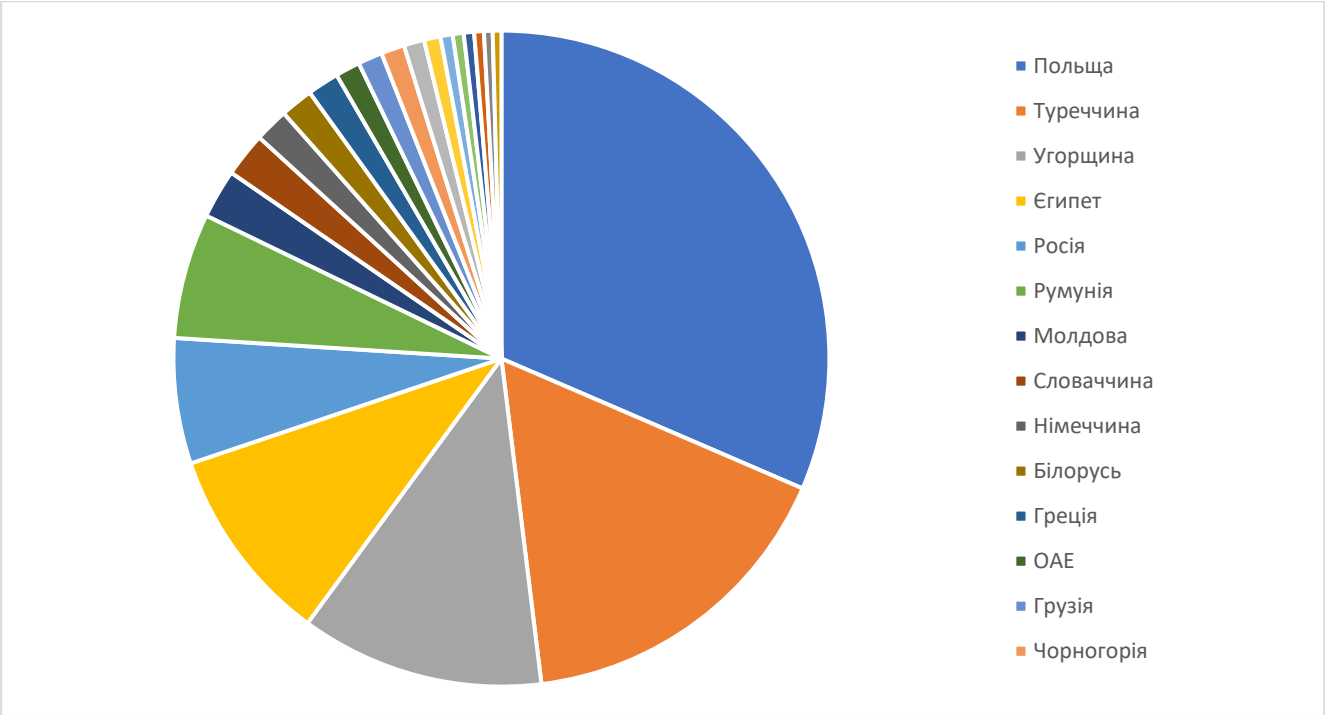


Рисунок 1.2 – Співвідношення чисельності туристів

Як видно з рис. 1.2 найбільш відвідуваною є Польща, яка прийняла в себе 3 301 511 туристів (наведені дані за 2021 рік до повномасштабного вторгнення в Україну)

1.2 Системи рекомендацій та їх використання в туризмі

Щоб ефективно розробити успішну систему рекомендацій місць призначення необхідно подолати кілька проблем, а саме:

1. Покращення процесу прийняття рішень

Одним із завдань СРМП є покращення процесу прийняття рішень туристом. Туристам важливо розуміти, як були визначені рекомендації, які генерує система. Щоб досягти цього, потрібне глибоке розуміння прийняття рішень туристами та розробка нових моделей для процесу пошуку інформації [7]. Розуміння процесу прийняття туристичних рішень привертає увагу як дослідника, так і практики.

2. Зменшити роботу користувачів і зберегти їх конфіденційність

Необхідно усунути невизначеності, які виникають на етапі пошуку інформації в процесі прийняття туристом рішення. Зокрема, слід виключити будь-

який вхід користувача, який є несуттєвим для процесу пошуку (що забезпечить конфіденційність користувача). Включення більшої кількості параметрів у систему може збільшити складність моделі, знизити ефективність СРМП і знизити рівень задоволеності користувача системою.

3. Підвищення ефективності рекомендацій

Багато існуючих ТРС оцінюють систему лише за показником точності, і багато з них не мають жодного методу оцінки [6]. Підвищити ефективність можна використовуючи визначення рівня точності класифікації. Підвищення ефективності рекомендацій під час процесу побудови моделі є складним завданням. Існує багато методів підвищення ефективності системи рекомендацій. В цій роботі увага зосереджена на дослідженні алгоритмів класифікації, параметрів оптимізації та комбінування класифікаторів. По-перше, необхідно провести дослідження алгоритмів множинної класифікації, оскільки деякі алгоритми краще підходять для наших наборів даних, ніж інші. Можна застосувати різні види методів перехресної перевірки, щоб переконатися, що модель не є надто складною та достатньо узагальненою для невидимих даних. По-друге, налаштування гіперпараметрів для алгоритмів класифікації є вирішальним процесом для підвищення точності прогнозування. Однак налаштування гіперпараметрів вважається дорогим і трудомістким процесом. Ці гіперпараметри відіграють важливу роль у прогнозуванні результатів, а метою є пошук оптимальних. По-третє, було доведено, що метод ансамблевого навчання дає кращі результати, оскільки цей метод об’єднує результати кількох базових класифікаторів [14]. Основна проблема на сьогодні полягає в тому, що невідомо, який метод комбінування дасть кращі прогнозні результати. Таким чином, нам потрібно провести дослідження, щоб порівняти результати двох типів методів ансамблевого навчання, включаючи методи, які поєднують кілька моделей подібного типу, і методи, які поєднують кілька моделей різних типів.

4. Підвищення рівня задоволеності користувачів

Ще одна проблема у розробці СРМП пов’язана з підвищенням рівня задоволеності користувачів системою, забезпечення зручного інтерфейсу та легкої взаємодії користувача з системою.

Системи рекомендацій, підмножина систем підтримки прийняття рішень, є інструментом, який може рекомендувати елемент на основі агрегування уподобань користувача [1]. Він надає цінну інформацію, яка допомагає користувачам приймати рішення на основі пріоритетів і проблем [13]. РС зазвичай застосовують свою методологію з трьох сфер: інформаційний пошук, взаємодія людина-комп’ютер та інтелектуальний аналіз даних [13]. Системи рекомендацій відіграють важливу роль на багатьох популярних веб-сайтах електронної комерції, таких як Netflix, Spotify, Pandora, Amazon і LinkedIn, а також на інших, пропонуючи користувачам елементи, зокрема фільми, музику, новини, статті, людей і URL-адреси [3]. Системи рекомендацій широко застосовуються в багатьох областях, але в даній роботі розглянуто їх застосування лише у сфері туризму, які називаються туристичними рекомендаційними системами (ТРС).

Характерною рисою туризму є прийняття рішень туристом або туристичним агентом: процес вибору місць призначення, пам’яток, заходів, готелів, ресторанів і послуг. Саме тому більшість досліджень зацікавлені саме у використанні туристичних рекомендаційних систем. Ґрунтуючись на введених користувачами даних, ТРС можуть:

1. Рекомендувати результати, на основі оцінки інтересу користувача.
2. Рекомендувати об’єкти, туристичні послуги або маршрути.
3. Ранжувати запропоновані напрямки в послідовності.
4. Запропонувати цілісний план поїздки.

На сьогодні уже існують системи, які допомагають у виборі не лише туристу, а й підтримують ведення бізнесу турагентів. Вони мають схожі характеристики та обмеження, але відрізняються вибором технологій, методами для покращення персоналізації, введенням даних, стилем взаємодії та методами рекомендацій.

Рекомендована система може складатися з кількох підсистем, наприклад оптимізаційні, статистичні та інтелектуальні підсистеми. Ці підсистеми

використовуються для пропонування, ранжирування або прогнозування елементів, таких як пункти призначення, пам’ятки, заходи та послуги на основі вимог користувачів, уподобань, жорстких і м’яких обмежень, таких як демографічна інформація користувача, кількість днів подорожі, бюджет подорожі та тип подорожі.

Як правило, перед або під час подорожі турист надає вхідні дані (наприклад, неявні, явні або обидва) до TPC, який потім створює профіль користувача та обчислює рекомендовані результати на основі профілю та різних баз даних. TPC може представляти результати багатьма способами, наприклад, використовуючи піктограми пункту призначення в інтерфейсі карти з маршрутом «точка-точка», маршрутом. Більшість TPC надають результати за допомогою просторових веб-сервісів і програмного інтерфейсу Google Maps (API).

Деякі TPC тепер можуть адаптувати свої результати до користувача, включаючи інформацію про контекст користувача, таку як місцезнаходження чи погода. Деякі TPC дозволяють користувачеві змінювати результати за допомогою відгуків користувачів або оцінок користувачів; тоді TPC можуть оновлювати профілі користувачів, щоб давати майбутні рекомендації [14].

У цьому дослідженні запропоновано розробити рекомендаційну систему, яка рекомендує туристам напрямки. Наша система прийняття рішень має дві основні відмінності порівняно з попередніми системами, які можна знайти в літературі: посилення системи за рахунок введення блоку машинного навчання та розробки зручного інтерфейсу користувача.

1.3 Використання сучасних інформаційних технологій при розробці систем рекомендацій

Сучасні системи рекомендацій можна значно підсилити за рахунок використання інформаційних технологій: бездротові сенсорні мережі, Інтернет речей, машинне навчання, Web 3.0, агентна технологія та багато інших (рис. 1.3).

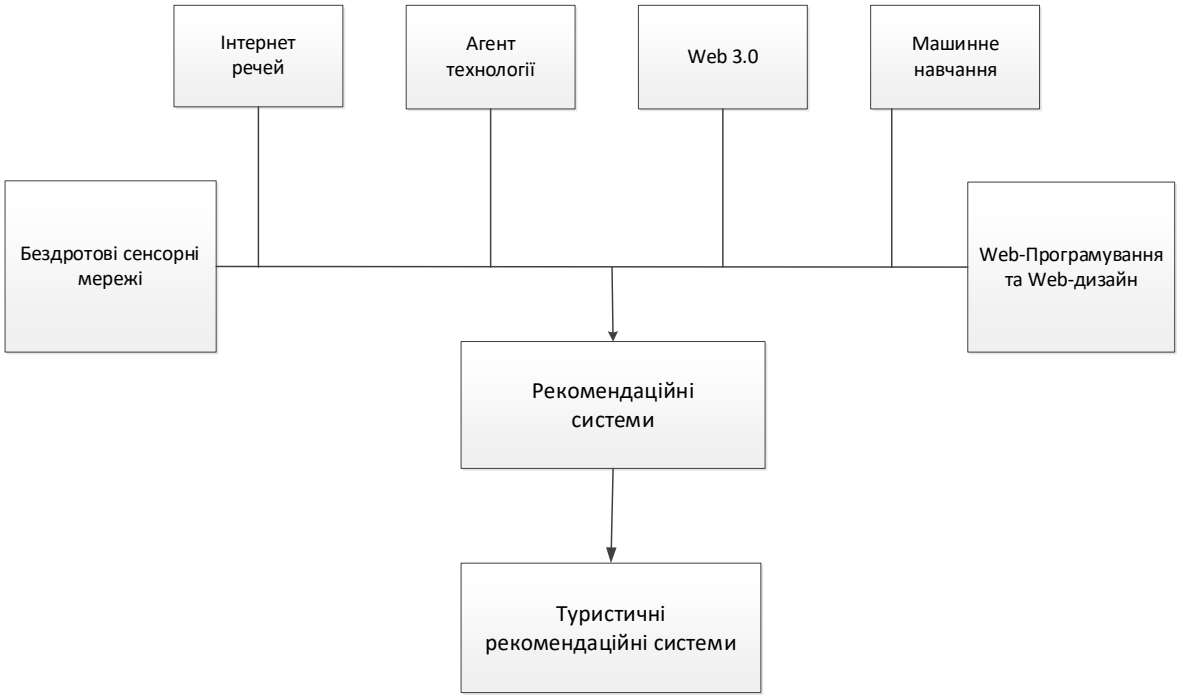


Рисунок 1.3 – Інформаційні технології у туристичних рекомендаційних системах

Бездротові сенсорні мережі вдосконалюють системи рекомендацій для туристів з точки зору усвідомлення контексту, рекомендацій у реальному часі, можливостей перепланування маршруту під час подорожі та адаптації до змін.

Глобальна система позиціонування (GPS) і географічні інформаційні системи (GIS) використовуються для отримання інформації про місцезнаходження користувачів, надання вказівок користувачам, виявлення друзів поблизу, розрахунку швидкості подорожі та виявлення прилеглих цікавих місць. Технології GPS і GIS допомагають користувачеві знаходити найкращі місця або маршрути як до, так і під час подорожі.

Багато ТРС розгортаються не лише як окремі програми на настільних комп’ютерах або платформах браузерів, але також підтримуються на мобільних пристроях через поширеність смартфонів із вбудованим GPS, компасами, акселерометрами та іншими датчиками. За допомогою мобільних додатків для рекомендацій можна використовувати такі параметри, як погода, рівень шуму та люди поблизу. Крім того, мережі зв’язку 3G, 4G, Wi-Fi, WiMAX і Bluetooth надають дослідникам більше можливостей і нові сучасні ресурси.

Проте є дві проблеми щодо впровадження інновацій для ТРС. По-перше, це використання контекстно-залежних рейтингів як підхід до спільної фільтрації, де туристи можуть завантажувати, переглядати та залишати коментарі через свої мобільні пристрої. По-друге, існує спроба запровадити інфраструктуру бездротової сенсорної мережі (WSN), щоб вирішити проблему забезпечення економічно ефективних засобів для віддаленого оновлення вмісту та підтримки виявлення близькості (розташування цікавих місць у сільській місцевості). Вхідні дані надходять із реєстрації на веб-сайті користувача, де вхідні змінні можуть включати стать, сімейний стан, вік, рівень освіти, категорії місць та улюблені види дозвілля за бажанням. WSN – це інновація, яка через відсутність розвиненої мережевої інфраструктури та високу вартість мобільних послуг у багатьох країнах призводить до того, що туристи здебільшого уникають використання 3G/Edge-з'єднань [7].

Інтернет речей (IoT) – сукупність технологій, які можуть відігравати важливу роль у туристичній індустрії. IoT відноситься до тенденції злиття фізичного світу зі світом інформації в загальному стані зв'язку, подібному до Інтернету. Наприклад, IoT об'єднує багато об'єктів, зацікавлених сторін, агентів і підсистем у їхніх бізнес-процесах. Таким чином, туристи тепер можуть генерувати, надсилати та отримувати дані за допомогою комунікаційних пристроїв за допомогою низки комунікаційних технологій, мережевих протоколів і типів даних без втручання людини.

Штучний інтелект (AI) – технологічні та наукові рішення, які включають в себе Data science і Machine learning і дозволяють інтелектуалізувати системи прийняття рішень, а саме: покращення процесу прийняття рішень, оптимізації, планування, кластеризації, представлення знань і планування.

До таких методів належать байєсовські мережі, іноді відомі як мережі переконань або ймовірнісно спрямовані ациклічні графічні моделі, є одним із найпопулярніших методів машинного навчання, які доречно використовувати для оцінки вподобань користувача на основі попередньої інформації. Байєсовські

мережі – це гібридна система рекомендацій, яка поєднує фільтрацію на основі вмісту та спільну фільтрацію.

Нечітка логіка також може бути використана для ТРС, з метою усунення невизначеностей, які оточують лінгвістичні оцінки, отримані від галузевих експертів і відгуків туристів.

Case-Based Reasoning (CBR), метод машинного навчання, надає рішення подібних проблем за допомогою чотирьох процесів: отримання, повторного використання, перегляду та збереження.

Альптекін і Бююкозкан [2] запропонували інтелектуальну систему планування туристичних напрямків, щоб допомогти туристичним агентствам зменшити навантаження. Система поєднує в собі CBR і багатокритеріальне прийняття рішень для підвищення точності системи, де обидва методи мають щось спільне з точки зору прийняття рішень. Проблеми цього дослідження включали інтеграцію цих двох методів прийняття рішень і розуміння того, як підвищити точність ТРС. Вимоги користувача, такі як тип туру (наприклад, активний, мандрівний, місто), кількість мандрівників, регіон, вид транспорту, тривалість туру, сезон, тип розміщення та рейтинг (тобто кількість зірок) є параметрами для ТРС. Результатом цього ТРС є план подорожі з указаною ціною. Перевагами системи є те, що надійність отриманих результатів і структура може бути адаптована відповідно до інших областей застосування. Основним недоліком цієї системи є функція адаптації, яка значною мірою залежить від досвіду туристичних агентств. Наприклад, коли турист створює нову справу, її неможливо вставити безпосередньо в базу даних; скоріше його має оцінити туристичне агентство або спершу прийняти турист (тобто фаза адаптації виконується офлайн або вручну). Іншим недоліком є проблема холодного запуску (тобто система не має достатньо інформації, щоб зробити будь-які висновки про користувачів), оскільки ця ТРС вимагає багато часу для збору даних і передачі їх до бази даних.

Генетичний алгоритм — це евристика пошуку, яка імітує процес природної еволюції.

Існує багато методів штучного інтелекту, які використовують механізми рекомендацій за межами сфери ТРС. Назвемо декілька:

1. Методи матричної факторизації, призначені для рекомендувача та використані в підході спільної фільтрації в рекомендаціях фільмів, використовуючи набір даних Netflix.

2. Система рекомендацій з урахуванням вартості, спрямовану на надання рекомендацій щодо подорожей з урахуванням вартості. Система прогнозує користувачам туристичні пакети на основі вартості поїздки та інтересу туриста. Система використовує дані туристичного туру, зібрані від туристичної компанії, використовуючи процеси Гауса для розробки моделі, і оцінює систему за допомогою метрики.

3. Система рекомендацій на основі комунальних послуг для прогнозування функцій корисності споживачів та їх платоспроможності. Система розроблена на основі введення лише порядкових атрибутів, і системи, які використовують методи спільної фільтрації, можуть отримати вигоду від їх підходу.

4. Система рекомендацій, яка пропонує оптимальні запитання для використання на веб-сайті як введення користувача [4].

Хоча ці дослідження представляють інтерес, їх системні цілі зосереджені на точності прогнозування, а не на проблемі застосування в туризмі. Тому доречним є розробити систему, яка не лише зосередилися на точності прогнозування, але й зосередилися на прозорості та інтерпретованості моделей.

Дерево рішень — це ієрархічна модель, вона надає правила прийняття рішень, які полегшують розуміння процесу прийняття рішень.

У сфері ТРС більшість розроблених моделей вважаються чорними ящиками. Тому в роботі запропоновано підхід, який розроблений за принципом білого ящика, рекомендацій щодо місця призначення порівняно з іншими трьома системами полягає в тому, що використовується гібридний підхід, який складається з підходів фільтрації на основі вмісту, співпраці та знань.

Онтологія та Web 3.0

Метою Web 3.0 є ефективний обмін даними та обробка інформації в автоматичному та ручному режимах шляхом просування загальних протоколів обміну та форматів даних. Щоб представити знання в сфері туризму, зазвичай використовується технологія, яка називається онтологією. Онтологія – це метод формалізованого представлення знань в певній предметній області, який використовується в інформатиці та інформаційних науках. Цей метод розглядає зв'язки всередині бази знань, а також відіграє важливу роль у структурі семантичної мережі. Технологія семантичної мережі та онтологія допомагають дослідникам інтегрувати різнорідну онлайн-інформацію.

Агентна технологія має багато переваг при моделюванні складних проблем реального світу [8]. Багато персоналізованих систем туристичних рекомендацій використовують саме цю технологію. Багатоагентна система (MAS) складається з агентів, які взаємодіють один з одним у середовищі. Кожен агент має власну мету і намагається максимізувати ресурси системи. MAS є багатообіцяючими інструментами для моделювання проблем організації або проблем реального світу, де люди повинні приймати рішення як група [11]. Деякі агенти в системі ідентифікуються як інтелектуальні агенти, оскільки вони можуть приймати рішення, оптимізувати, планувати та вирішувати складні проблеми.

Відповідно до рис.1.4, використання багатоагентної системи має багато переваг для розподіленої системи, оскільки існує агент, що працює на мобільному пристрої, агент-посередник, який працює як фасилітатор між агентом користувача та агентом активності для обробки зв'язку між ними. та інший агент, відповідальний за обслуговування баз даних, щоб зменшити перевантаження сервера тощо. Крім того, здатність адаптувати, налаштовувати, додавати та видаляти агентів здається підходящою концепцією для модульного проектування при моделюванні розподіленої системи та реальних проблем. Крім того, у системи є високий ступінь адаптивної здатності, так що система може налаштовувати план, що грунтується на новому місці розташування користувача на момент виконання. Відгуки користувачів базуються як на явних (тобто підхід до рейтингів), так і на

неявних (тобто моніторинг дій шляхом аналізу часу, який користувач проводить на веб-сторінці, і посилань, за якими користувач переходить тощо) факторів.

Рисунок1.4 – Огляд архітектури системи на основі агентних технологій

Web-програмування та web-дизайн – технології, які дозволяють розробити веб-сайти, які будуть інтерактивними, інформативними та привабливими.

Щоб задовольнити це очікування, багато персоналізованих туристичних рекомендаційних систем використовують веб-програмування AJAX, яке поєднує кілька технологій, таких як HTML, JavaScript, XML і модель документ-об’єкт, щоб створити відчуття взаємодії між користувачем і web-додатком. Web-дизайн є одним із найважливіших технологічних інновацій для індустрії туризму. Крім того,

функції доступності для людей з обмеженими можливостями та людей похилого віку слід розглядати як корисну функцію для інтерактивного web-сайту.

1.4 Аналіз туристичних рекомендаційних систем

Багато нещодавніх ТРС зосереджені на рекомендаціях місць призначення разом з інтеграцією певних туристичних послуг, таких як готелі та ресторани. Результати більшості систем базуються на маршруті. Останнім часом дослідники розширили свою увагу, включивши рекомендації маршрутів і розв'язання проблем, пов'язаних із проектуванням подорожей/маршрутів. Багато ТРС пропонують цілісний план поїздки, головним чином зосереджуючись на конкретному вмісті. Згідно з літературою, ТРС можна класифікувати на основі послуг електронного туризму, які вони надають, включаючи рекомендації щодо місць призначення, рекомендації щодо туристичних послуг, рекомендації щодо маршрутів та рекомендації щодо планування подорожей/маршруту.

Навіть найпростіші ТРС перераховують пункти призначення відповідно до певних обмежень введення, наданих користувачами. Деякі з них враховують контекстну інформацію. Системи рекомендацій місць призначення рухаються до точки, коли вони зможуть ранжувати важливість пунктів призначення та прогнозувати придатність пункту призначення для користувача. Деякі системи використовували теорію прийняття рішень, щоб краще зрозуміти, як туристи вибирають бажані напрямки, щоб підвищити точність прогнозування.

INTRIGUE пропонує як веб-платформи, так і мобільні (кишенькові пристрої) платформи для міста Турин, Італія. Система рекомендує POI (тобто місця огляду визначних пам'яток) і маршрути, беручи до уваги переваги різнорідних туристичних груп (наприклад, сім'ї з літніми людьми або дітьми), оскільки це є однією з проблем поточного дизайну ТРС. Ця рекомендаційна система враховує багато обмежень користувача, як-от кількість днів, час прибуття/від'їзду, початкове та кінцеве місце розташування та бажаний час відвідування. Механізм рекомендацій цієї системи значною мірою покладається на методи моделювання

користувача та гіпермедіа. Ця система також підтримує планування турів як до, так і під час подорожі, що є ще одним викликом для розробки рекомендаційної системи.

PSiS – це мобільний TPC, який надає рекомендації щодо POI, зосереджуючись на контексті користувача (наприклад, місцезнаходження, час, швидкість, напрямок, погода) і вподобаннях користувача. Система має можливість динамічно адаптуватися до рекомендованого туру; наприклад, вона може створити новий план поїздки, коли користувач випереджає графік. Іншим завданням цього TPC є імплантація проміжного програмного забезпечення, яке знаходиться на сервері. Він синхронізує дані між веб-програмою та мобільною програмою. Іншою цікавою функцією є архітектурний тег, який може рекомендувати POI відповідно до того, відкритий чи закритий пункт призначення та який варто відвідати. Додатковою функцією є система відстеження, яка економить час.

SPETA використовує переваги технологій Web 3.0, інтегруючи соціальні мережі, семантичну мережу та контекстну інформацію у мобільну рекомендаційну систему. Система має на меті рекомендувати туристичні послуги, такі як атракціони чи ресторани, туристам, які вперше відвідують цей регіон. TPC зосереджується на зіставленні, пошуку та фільтрації елементів із знань, отриманих за допомогою онтології (тобто соціальної та геолокаційної інформації). Система вимагає введення даних від користувача, щоб надавати рекомендації. Вхідні дані містять уподобання користувача (типи їжі та музики), контекстну інформацію користувача (погода, час, місцезнаходження) та інші дані, такі як швидкість і напрямок. Система також включає час відкриття та закриття та дати атракціонів.

SigTur/E-Destination – це веб-платформа TPC для планування подорожей, яка рекомендує види дозвілля в Таррагоні, Іспанія. Система враховує багато різних видів введення, як явно, так і неявно. Користувач повинен чітко ввести мотивацію подорожі, демографічну інформацію користувача (наприклад, країну походження), бюджет подорожі, склад групи, необхідний пункт призначення, тип проживання та дати подорожі (дати початку та кінця) через веб-інтерфейс. Коли користувач відповідає (тобто додає або видаляє інформацію) на результати рекомендацій,

система рекомендацій сприймає це як неявні вхідні дані для врахування в майбутніх рекомендаціях. Переваги цієї системи полягають у її гібридному рекомендаційному підході та методі прогнозування, який аналізує величезний набір даних.

Otium – це персоналізована система планування подорожей, яка планує дозвілля для туристів. Ця система покладається на методологію веб-вилучення для отримання інформації для своєї бази даних. Він використовує інтерактивний веб-інтерфейс, щоб користувач міг налаштувати створений розклад відповідно до своїх уподобань. У системі є два методи введення. Спочатку турист вказує максимальний бюджет і зону подорожі (місто/провінція). Крім того, близькість, ціна, час, профіль і різноманітність є параметрами, які необхідні для розрахунку плану поїздки в рекомендувачі через веб-інтерфейс. Цей метод є перевагою при роботі з джерелами веб-інформації. Однак оболонка може аналізувати лише файл HTML. Він повинен йти в ногу з файлом конфігурації, щоб мати можливість адаптуватися до змін у структурі файлу HTML. Крім того, він може витягувати лише атрибути події. Цій ТРС бракує багатьох важливих функцій, наприклад функція транспортування, за допомогою якої користувач може шукати вид транспорту для вибору під час подорожі. Іншим недоліком є навігаційна система, оскільки вона може використовувати зібрані геопозиції для нанесення маршруту або розташування за допомогою Google Maps.

SAMAP – це ТРС, призначений для допомоги туристам у плануванні подорожі на основі історії користувачів та інших факторів. Він зосереджений на задачі командного орієнтування з часовими вікнами і рекомендаціях щодо діяльності. SAMAP базується на мультиагентній системі та призначений для роботи на мобільних пристроях. Системні вхідні дані включають налаштування користувача, особисту інформацію та контекст користувача. Транспорт (наприклад, автобус, таксі, пішки) та інформація про навколишнє середовище (наприклад, затори, тип вулиці) також враховуються. Система рекомендує план подорожі зі списком заходів для відвідувачів і пропонований маршрут, який починається від однієї POI і потім веде користувача до іншої.

e-Tourism – це гібридний TPC, який зіставляє демографічні дані та вподобання користувачів із базою даних місць призначення для створення плану дозвілля зі списком рекомендованих видів дозвілля в Іспанії. Для опису туристичної діяльності використовується таксономія, набір понять. TPC використовує планування штучного інтелекту для створення реалістичних планів діяльності, включаючи години роботи, пріоритети, тривалість відвідування та корисність як обмеження. Система є адаптивною, використовує систему оцінювання після входу користувача для отримання зворотного зв'язку з метою покращення профілю користувача.

1.5 Постановка завдання магістерського дослідження

Ця магістерська робота пропонує інноваційну систему рекомендацій місць призначення (СРМП) для відповіді на виклики сьогодення. Запропонована СРМП вважається системою рекомендацій призначення на основі моделі. Контрольований процес машинного навчання, який виконується в автономному режимі, включає збір даних, попередню обробку даних, аналіз даних та інтерпретацію результатів.

Основною метою цього дослідження є покращення процесу вибору туристичних маршрутів за допомогою інформаційної технології на основі рекомендаційних систем з використанням методів машинного навчання.

Для досягнення зазначеної основної мети були встановлені наступні завдання:

1. Проаналізувати існуючі рекомендаційні системи, зокрема в сфері туризму. Визначити основні недоліки систем.
2. Визначення обмежень для системи.
3. Визначення вхідних даних та їх опрацювання.
4. Вибір методів машинного навчання та розробка алгоритму для їх ефективного застосування.
5. Розробка методики рекомендації туристичного маршруту.

6. Створення прототипу запропонованої інформаційної технології.

7. Визначення напрямків подальшого удосконалення проведеного дослідження.

Більшість ТРС зосереджуються на рекомендаціях пунктів призначення, маршрутів і реалістичного планування подорожей/маршрутів. Крім того, сучасні інформаційні технології надають дослідникам нові можливості для проектування та розробки ТРС, які є більш інтелектуальними, інтерактивними, адаптивними та автоматизованими, а також здатними запропонувати вищий рівень задоволеності користувачів і користувацького досвіду, ніж будь-коли раніше. Подальші ТРС мають спиратися на існуючі основи прийняття рішень, щоб бути більш ефективними та менш нав’язливими.

Це дослідження має на меті зробити внесок у розробку вдосконаленої системи прийняття рішення, оскільки в попередніх системах бракує як технічних методів, таких як точність рекомендацій і оцінка, так і практичних аспектів, таких як задоволеність користувачів. Саме тому слід розробити систему, яка розуміє як обрати місце призначення туриста шляхом розробки моделей вибору місця призначення з використанням як кількісних, так і якісних підходів, а також підвищення рівня задоволеності користувачів за допомогою машинного навчання та методів веб-технологій.

2 ВИБІР ДАНИХ ТА ПОБУДОВА МОДЕЛІ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ

2.1 Введення системи обмежень для користувачів

Технології бездротової сенсорної мережі, такі як GPS і RFID, можуть отримувати контекстну інформацію, наприклад поточне місцезнаходження, як параметр. Маршрутні ТРС дають рекомендації від точки до точки за допомогою мультимодельних транспортних послуг Крім того, існує ТРС, які надають інформацію в режимі реального часу туристам, щоб зменшити затори та уникнути довгих черг у туристичних точках.

Планування поїздки є складним завданням; наприклад, туристи зазвичай мають конкретні вимоги та потреби, такі як кількість днів подорожі, кількість мандрівників, бюджет, необхідні напрямки, дні, коли заклади працюють, і початкові місця. Системи планування подорожей/рекомендації маршруту враховують ці вимоги користувача під час визначення порядку місць призначення в маршруті. Крім того, ці системи можуть створювати новий план/маршрут для мандрівника у відповідь на зміни, що відбуваються під час подорожі. Наприклад, якщо мандрівник запізнюється, система може змінити графік призначення.

Хоча ТРС охоплюють багато різних аспектів туристичних послуг, мало хто зосереджується на проблемі планування поїздки чи розкладу, оскільки це складна проблема, яка вимагає від ТРС генерувати автоматизований оптимальний план подорожі (тобто найбільш реалістичний план подорожі) для користувача на основі на багатьох обмеженнях. Ця проблема була названа проблемою проектування туристичного маршруту – нагадує класичну проблему комівояжера у дослідженні операцій. Однак проблема ТРС пов’язана з мінімізацією часу подорожі або відстані подорожі; найпростіший можна змоделювати як задачу орієнтування, де набір вершин містить задані точки інтересу, кожна з яких має оцінку (наприклад,

задоволеність користувача), а метою є створення найкращого шлях, щоб максимізувати загальний бал (час або бюджет) для кожної з вершин.

Рекомендація майже оптимального або реалістичного маршруту поїздки є серйозною проблемою, тому наступні обмеження користувача та контекстні обмеження можна додати до TPC для створення більш реалістичних і ефективних рекомендованих планів поїздок. Це робиться, щоб задовольнити вимоги та переваги користувачів.

TPC надають варіанти під час вибору місць призначення та послуг, беручи до уваги жорсткі обмеження користувача, включаючи контекстну інформацію, вимоги, уподобання, інтереси, демографічні дані та інформацію про пункт призначення. Майбутні TPC мають надати мандрівнику ще більше можливостей, щоб змусити систему збирати інформацію про пункти призначення, які хоче відвідати, залежно від його потреб. Наприклад, деякі туристи не хочуть відвідувати більше певної кількості місць призначення на день або місць призначення, які вони вже відвідали під час попередньої подорожі. Оскільки більшість користувачів зважають на бюджет, бюджет поїздки має включати ліміти на транспортні збори, плату за вхід на подію та рахунки за готель та ресторан. Крім того, систематично слід враховувати перерви на обід або вечерю, перерви на каву та короткі перерви протягом дня. Надаючи системі часові рамки для таких перерв, система зможе знайти інші пов’язані пункти призначення чи послуги з робочими годинами, які відповідають вказаному користувачем доступному часі.

Крім того, слід враховувати кількість днів подорожі та проблеми з доступністю (наприклад, порушення зору або слуху, рухові порушення). Можна побачити, що майбутні TPC, які стурбовані реалістичним планом поїздки, повинні досліджувати механізми розвідки, які можуть запускати оновлення маршруту, коли контекстна інформація змінюється.

До TPC можна додати м’які обмеження. Наприклад, TPC, який рекомендує ресторани, можна запрограмувати на включення часу прийому їжі, типу їжі (китайська, тайська чи японська) і цінового діапазону (низький–високий). Завдяки цим м’яким введенням TPC може рекомендувати ресторани з годинами роботи та

діапазоном цін, які відповідають критеріям вибору користувача. Для ТРС, який рекомендує готелі, також можна додати м’які обмеження, такі як тип готелю, діапазон цін і зручності. Варіанти транспортування повинні базуватися на моделі з декількома варіантами та деяких інших аспектах щодо транспортних послуг (наприклад, транспортні збори). Щодо контекстної інформації; слід враховувати погоду, прогноз руху та поточну дату/час, щоб відповідати датам/часам роботи пункту призначення.

Існує простір для додаткових досліджень систем рекомендацій на основі обмежень і контексту не тільки в сфері туризму, але й щодо інших програм, включаючи навігацію по картах, управління автопарком, інформацію про погоду, допомогу на дорозі та персональні служби визначення місця розташування.

Рекомендувати майже оптимальний або реалістичний маршрут подорожі, який включає обмеження користувача та контексту, щоб задовольнити вимоги та переваги користувача, є ще одним викликом.

Інтеграція різнорідної онлайн-інформації про подорожі є серйозною проблемою для ТРС. ТРС передбачає збір великих обсягів інформації від різних постачальників інформації або туристичних послуг з різними або навіть унікальними типами категорій або вмісту в різних форматах, включаючи неструктурні дані. Щоб вирішити цю проблему, застосовуються методи вилучення інформації, такі як вилучення веб-сканерів, семантичні технології та технології Web 2.0.

Традиційна система керування реляційною базою даних матиме труднощі з керуванням великими обсягами та складним характером даних, що використовуються в ТРС, включаючи геопросторові дані, постійні та численні оновлення користувача, враховуючи проблеми з доступністю даних та масштабованістю.

Групові системи рекомендацій викликають труднощі, тому що не тільки групи туристів мають різні індивідуальні переваги, але вони також повинні зважати на переваги інших членів групи. Важко порекомендувати маршрут для групи, який би оптимально відповідав різним індивідуальним інтересам.

Ще однією вимогою до рекомендаційних систем є наявність інтерактивних web-програм, що швидко реагують. При перегляді туристичних web-сайтів, користувачі очікують, що вони будуть інтерактивними, інформативними та привабливими. Щоб відповідати цим очікуванням, багато персоналізованих туристичних рекомендаційних систем використовували веб-програмування AJAX, яке поєднує кілька технологій, таких як HTML, JavaScript, XML і об’єктні моделі документів, щоб створити відчуття взаємодії між користувачем і веб-додатком.

Прийняття туристами рішень та обробка інформації має бути за допомогою людиноцентричного підходу.

Фактори, які впливають на пошук подорожей і процеси прийняття мандрівниками рішень:

1. Особисті характеристики мандрівника (наприклад, соціально-демографічні показники, знання, особистість, цінності, ставлення, когнітивний стиль, стиль прийняття рішень, стиль відпустки). Основними факторами, які впливають на рішення споживача при покупці продукту або послуги, є джерела інформації про цей продукт або послугу. Крім того, індивідуальні демографічні показники можуть впливати на поведінку в пошуку інформації.

2. Характеристики подорожі (наприклад, мета подорожі, тривалість поїздки, відстань подорожі, група подорожей, мобільність подорожі)

3. Фактори навколишнього середовища (джерела інформації, культура, сім'я, спосіб життя та особливості місця призначення) та фактори індивідуальних рис (мотивація, особистість і минулий досвід)

4. Особистий досвід – є найважливішим чинником у підвищенні обізнаності про напрямок.

Збільшення використання мобільних телефонів і нові розробки мобільних комп’ютерів і комунікаційних мереж (наприклад, GPS, Wi-Fi) пропонують найсучасніші вдосконалення систем рекомендацій у сфері туризму.

Більшість попередніх ТРС підтримували лише окремих туристів і зосереджувалися на оцінках під час вибору місця призначення, заходів, пам’яток і туристичних послуг (наприклад, ресторани, готелі, транспорт) на основі вподобань

та інтересів користувача. Що стосується технічних аспектів, ці ТРС забезпечують лише фільтрацію, сортування та основні механізми зіставлення між елементами та жорсткими обмеженнями користувача.

Слід зазначити, що сучасні ІТ-технології надають дослідникам нові можливості для розробки та реалізації ТРС, яка є більш інтелектуальною, інтерактивною, адаптивною та автоматизованою, такою, яка підтримує вищий рівень задоволеності користувачів, ніж будь-коли раніше. Тому якщо узагальнити результати проведених досліджень, можна сказати, що сучасні ТРС повинні володіти наступними властивостями:

1. Покращений процес прийняття туристичних рішень.
2. Зменшити зусилля користувача.
3. Продуктивність, швидкість, точність рекомендацій і точність СРМП.
4. Інтелектуальний інтерфейс користувача або веб-сайт.
5. Інтеграція різнорідної інформації.
6. Побудова цілісного плану подорожі.
7. Оптимальні рекомендації для групи туристів.
8. Високоадаптивна система.
9. Забезпечення конфіденційності користувачів.

Поточні ТРС починають збирати більше інформації від користувача, але передача певної інформації може вважатися конфіденційною.

2.2 Отримання необхідного набору даних та їх обробка

В запропонованому дослідженні використано п’ять наборів факторів, які впливають на вибір туристичних напрямків – фактори мотивації: включаючи самореалізацію, відпочинок, новизну, пригоди, досвід навчання, стосунки, соціальний статус і покупки.

Отже загалом фактори, які аналізувалися, це:

1. Характеристики подорожі: тривалість подорожі, мета, склад поїздки тощо. Характеристики туриста включають змінні психологічного, когнітивного та

соціально-економічного статусу, які впливають на процес вибору туристом місця призначення. Дані факти є найважливішими для алгоритму.

2. Сума витрат на поїздку: загальні витрати, які турист виділяє на поїздку та ділиться на кілька частин (тобто покупки, проживання тощо).

3. Бажані види діяльності: змінні психологічного, когнітивного та соціально-економічного статусу, які впливають на процес вибору туристом місця призначення.

4. Мета подорожі. Ця змінна описує причини, чому турист обирає відвідати певний пункт призначення.

5. Задоволення туристів характеристиками подорожі (ціна, харчування тощо). Змінні використовують на етапі інтерпретації результатів. Ці змінні мають діапазон значень від 1 до 5.

6. Демографічна інформація про туристів (вік, стать, дохід тощо).

Було розроблено пілотне дослідження. Пілотне дослідження мало на меті вивчити користувачів і дизайн запропонованого підходу. Цілі пілотного дослідження полягали в тому, щоб перевірити відповідність вхідних параметрів і вихідних даних запропонованої туристичної рекомендаційної системи, щоб зібрати вимоги користувачів, перевірити дослідницькі питання/проблеми та визначити будь-які потенційно нові.

У пілотному дослідженні використовувалася анкета з 20 відкритими запитаннями, яка проводилася протягом однієї години. Її отримали п’ять обраних учасників. Пілотне дослідження було проведено таким чином:

1. Знайомство учасника.

2. Ознайомлення з системою персоналізованих рекомендацій.

3. Відкриті питання.

Під час пілотного дослідження було виявили, що Інтернет є основним джерелом інформації для користувачів під час планування подорожі. Було також визначено, що доступ до персоналізованої системи рекомендацій буде оптимальною метою користувача.

Користувачі вважають, що системи рекомендацій допомагають людям, коли вони стикаються зі складними завданнями, і що вони повинні бути надзвичайно комплексними, як у цілісному плані. Попередня інформація, зібрана на основі досвіду туристів, зіграла важливу роль у розробці кращої системи допомоги користувачам у прийнятті рішень. Пілотне дослідження також виявило, що учасники хотіли програмне забезпечення, яке містить найновішу інформацію про визначні місця. Що стосується системної платформи, то комплексна платформа має вирішальне значення для реалізації цієї послуги, а також ефективність взаємодії з користувачем і простота програмного забезпечення. Що стосується відповідних вхідних даних, які користувач готовий ввести в систему, користувачі, швидше за все, нададуть вхідні дані, які не містять приватних чи особистих даних, наприклад дати, бюджет тощо. В більшості люди не хотіли ділитися даними, які стосуються демографічної інформації про турист (вік, стать, дохід, тощо).

Що стосується самої системи, то користувачі хотіли б мати результат у вигляді графічного зображення та текстової інформації. Представлення результатів було дуже важливим і повинно бути легким для розуміння. Усі учасники зійшлися на думці, що найбільше користі від запропонованої системи отримають туристи. Користувачі воліли б використовувати систему до початку поїздки, але система, яка дозволяє користувачеві коригувати план під час поїздки, також вважалася важливою. Крім того, він має бути доступним як мобільний додаток для зручності користувача. Що стосується механізму зворотного зв’язку з користувачами, то функція масштабування та перегляду або поєднання обох видалася найбільш бажаною.

Отже, найважливішими характеристиками системи на думку користувачів має бути: конфіденційність користувачів, групові рекомендації, взаємодія користувача з системою, мобільність, інтеграція різнорідної інформації та прагнення до цілісного плану поїздки.

Запропонована структура рекомендацій місць призначення складається з п’яти підсистем, заснованих на процесі інтелектуального аналізу даних:

1. Отримання даних.

2. Попередня обробка даних.
3. Вибір ознак.
4. Класифікація та побудова моделі.
5. Інтерпретація результатів.

Після проведення анкетування було сформовано звіт даних. Потім зібрані дані були попередньо оброблені за наступним алгоритмом: очищення даних, перетворення даних і методи вибору ознак. Процес аналізу даних включає кілька алгоритмів класифікації. В роботі запропоновано використовувати такі методи як дерево рішень, метод опорних векторів та багатошаровий персептрон, які служать класифікаторами та використовуються для розробки оптимальних моделей вибору місця призначення, а також правил прийняття рішень. Щоб покращити ефективність рекомендацій, окремі класифікатори були об’єднані за допомогою кількох методів комбінування. Запропонована система була оцінена за допомогою кількох вимірювань, наприклад матриця точності. Правила прийняття рішень були передані механізму інтерфейсу користувача для створення веб-інтерфейсу користувача на основі заданих моделей.

Запропонована структура використовує п’ять факторів як вхідні змінні, які були отримані з анкети. Потім вони були використані як вхідні дані для визначення класифікації бажаних місць призначення туриста. Потенційні вхідні дані включали характеристики подорожей, поведінку туристів, поведінку туристів щодо витрат, мотивацію подорожей та демографічну інформацію туристів. Коефіцієнти задоволеності користувачів були використані на етапі представлення результатів

Дані реального світу неповні, зашумлені та суперечливі. Так користувачі можуть надавати неправильні дані, задля збереження своєї особистої інформації. А якісне прогнозування вимагає точних, достовірних даних. Тому перш ніж використовувати дані необхідно їх обробити шляхом інтеграції, очищення, трансформації та скорочення даних.

Попередня обробка даних – аналіз відсутніх значень, виявлення або видалення невідповідностей – є одним із найважливіших компонентів попередньої обробки даних. Очищення даних для цієї роботи складалося з шести кроків.

Перший крок передбачав виправлення невідповідностей у даних шляхом вибору лише релевантних вхідних даних та використання вторинних даних сфери туризму, взятих з огляду літератури. Метою другого кроку було видалити випадки та змінні з багатьма відсутніми значеннями. Третій крок мав на меті згладити зашумлені дані шляхом видалення будь-яких екстремальних значень. Наступний крок передбачав зменшення кількості значень безперервних ознак за допомогою простої техніки групування. Деякі характеристики потребували нормалізації, агрегування та узагальнення.

Останній крок мав на меті зменшити розміри набору даних шляхом видалення зайвих і дублюючих функцій, які не додали потужності прогнозування. Наприклад, користувачеві потрібно ввести лише кілька відповідних вхідних даних, щоб отримати необхідні результати рекомендацій від системи.

Початковий вибір є першим кроком у процесі очищення даних. На цьому етапі знання, отримані в сферах туризму, використовуються для вибору змінних, які не пов’язані з класами випуску. Наприклад, змінні задоволеності, місце опитування, дата опитування, коментар та ідентифікатор опитування були виключені з набору даних.

Відсутні значення можуть значно вплинути на аналіз даних. З метою усунення цього недоліку було використано наступні правила для видалення відсутніх випадків і змінних:

1. Випадки, пов’язані з відсутністю даних для прогнозованих змінних, були видалені, щоб уникнути будь-яких штучних посилень у їх зв’язку з незалежними змінними.

2. Змінні, у яких відсутні принаймні 10 відсотків даних, були кандидатами на видалення.

3. Випадки, у яких бракувало понад 15 відсотків даних, були кандидатами на видалення.

Для змінних, які класифікуються як випадково відсутні, для заміни відсутніх значень використовувався метод імпутації. Цей етап було зроблено для оцінки відсутніх значень на основі дійсних значень інших змінних або випадків у вибірці.

Одним із найпопулярніших методів є підстановка середнього значення або моди. Переваги використання методу заміни середнього значення або моди полягають у тому, що його легко реалізувати та дозволяє надати повну інформацію для всіх випадків. Метод підстановки середнього та моди найкраще використовувати, коли змінна має відносно низький рівень відсутніх даних.

Зазвичай в наборі даних з'являються екстремальні значення. Їх потрібно ідентифікувати та видалити, щоб зменшити дисперсію моделей.

При обробці даних можна використати методи дискретизації та нормалізації. Більшість алгоритмів не працюють добре для неперервних змінних, саме тому, їх потрібно перетворити на дискретні змінні. Неперервні змінні, такі як поведінка витрат, містять багато екстремальних значень. Але більш важливим є діапазон значень для кожної неперервної змінної.

У цьому дослідженні було застосовано два методи дискретизації. Перший метод дискретизації називається простим групуванням. Він ділить діапазон на N інтервалів однакового розміру. Нехай A і B – мінімальне і максимальне значення змінної; тоді ширина (W) інтервалу визначається за наступною формулою:

$$W = \frac{(B - A)}{N} \tag{2.1}$$

Другий метод дискретизації використовується для сортування даних і розділення їх на проміжки однакових розмірів; потім кожен проміжок згладжується з використанням середніх сум. Третій метод групування полягає в залученні експерта в домені, який вручну встановлював кількість категорій. Останній метод групування застосовується для обробки неперервних змінних, як описано в рівнянні нижче, де вибір значення для змінної *alpha* матиме вплив на процес вибору ознак, і це може бути обчислюється як:

$$x = mean \pm alpha \times std \tag{2.2}$$

Таблиця 2.1 – Приклад дискретизації щодо річного доходу домогосподарства

Діапазон, грн	Опис	Мітка
0	Дуже низький дохід	1
1000-3000	Низький дохід	2
3000-10000	Низький середній дохід	3
10000-20000	Середній дохід	4
20000-50000	Дохід вище середнього	5
50000-100000	Високий дохід	5
100000 і вище	Дуже високий дохід	6

Основною метою цього процесу було покращення продуктивності алгоритмів інтелектуального аналізу даних. Було застосовано три методи нормалізації даних: мінімально-максимальна нормалізація, нормалізація z-показника та нормалізація за допомогою експерта домену. Однак обраний метод залежить від обраного класифікатора. Наприклад, нормалізація min-max і нормалізація z-показника особливо корисні для класифікації алгоритмів, що включають опорні векторні машинні нейронні мережі, такі як класифікація найближчих сусідів. Однак вони можуть бути не дуже корисними при використанні в дереві рішень як моделі класифікації. Це може допомогти підвищити точність і простоту моделі дерева, але може спричинити труднощі з візуалізацією даних.

Мінімально-максимальна нормалізація виконується для лінійного перетворення даних до певних значень, зазвичай 0 і 1 або -1 і 1. Мінімально-максимальна нормалізація визначається за наступною формулою:

$$Normalized(f_i) = \frac{(f - F_{min})}{F_{max} - F_{\min}} \tag{2.3}$$

Нормалізація Z-показника виконує лінійне перетворення даних із використанням середнього значення($\bar{f}$) та стандартного відхилення (s). Нормалізація Z-показника визначається як:

$$Normalized(f) = \frac{\left(f - \bar{f}\right)}{s} \tag{2.4}$$

Що стосується третього методу, дані масштабуються до певного діапазону на основі знань експерта домену. Наприклад, змінна, яка описує «країну користувача», може містити 16 категорій/країн. Отже, дані у змінній можна масштабувати, як показано в таблиці 2.2.

Таблиця 2.2 – Нормалізація даних за допомогою експертних знань

Тип країни	Країни	
Розвинена	США, Японія, Франція, Німеччина	1
Країна, що розвивається	Україна, Китай, Малайзія	2
Нерозвинена	Лаос	3

Вибір функції є важливим кроком у попередній обробці даних перед переходом до процесу аналізу даних. Він передбачає вибір підмножини релевантних ознак для побудови класифікаційних моделей шляхом видалення нерелевантних і надлишкових ознак. Оптимальний вибір функцій надає багато переваг для покращення продуктивності алгоритму машинного навчання, зниження вартості зберігання даних тощо.

Вибір функцій є необхідним оскільки дозволяє краще зрозуміти, які змінні відіграють важливу роль, покращити продуктивність рекомендацій, зменшити кількість необхідних вводів користувача та підвищити ефективність моделі класифікації. Незалежна змінна, яка не пов’язана із залежною змінною, відома як нерелевантна ознака, тоді як незалежна змінна, яка не є корисною, відома як надлишкова характеристика, і її потрібно видалити перед побудовою моделі. Існує три типи методів вибору функцій, включаючи фільтри, згортки та гібридні методи. У методі фільтра змінні ранжуються та вибираються, передають алгоритму класифікації, який буде використовуватися. У методі згортки змінні відбираються з урахуванням алгоритму класифікації. Останній — гібридний метод, у якому

змінні спочатку вибираються за допомогою методу фільтра, а потім — методу згортки.

Вибірка є основною характеристикою, яка використовується в інтелектуальному аналізі даних або машинному навчанні для отримання підмножини набору даних. В дослідженні використовувалися вибірки з метою навчання, перевірки та тестування наборів даних для моделі. Навчальний набір даних використовувався для побудови моделі, а тестовий набір даних використовувався для оцінки моделі. Реальні дані є незбалансованими, тому для них існує багато стратегій вибірки, які були розроблені дослідниками для обробки незбалансованих даних, таких як недостатня вибірка, надмірна вибірка та штучна надмірна вибірка. В роботі використовувалася стратифіковану вибірка, щоб зменшити помилки вибірки. Стратифікована вибірка є найбільш прийнятним методом для процесу відбору моделі [19]. У стратифікаційній вибірці розділений набір даних містить однакові пропорції вихідних класів.

2.3 Вибір методу машинного навчання

В магістерській роботі було досліджено три традиційні алгоритми класифікації – дерево рішень, метод опорних векторів та багатошаровий персептрон. У наборі вхідних даних $D = (x_i, y_i)$, $i = 1,\ldots,n$, x складається з вибраних функцій із попереднього етапу, а y є пунктом призначення, пов’язаним з x, де $y \in \{c_1,\ldots c_n\}$ для n пунктів призначення. Набір вхідних даних D розділений на дві частини: навчальним та тестувальним набором. Навчальний набір використовується для навчання моделі, а тестовий набір використовується для оцінки ефективності класифікації навченої моделі. У побудові моделі є два основні процеси: процес вибору моделі та процес оцінки моделі. У процесі вибору моделі навчальний набір використовується для побудови моделі, а гіперпараметри класифікатора необхідно налаштувати, щоб отримати оптимізовану модель, як правило, через перехресну перевірку, визначену наступним чином:

$$CV(\theta) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in F_k}\left(y_i - \hat{f}_\theta^{-k}(x_i)\right)^2 \quad (2.5)$$

Наступним кроком обирається значення параметра налаштування, яке мінімізує помилку *CV* за наступною формулою:

$$\widehat{\theta} = \arg\min_{\theta\in\{\theta_1,\ldots,\theta\}} CV(\theta) \quad (2.6)$$

У процесі оцінки моделі для оцінки значення точності прогнозу використовується крос-валідація. Іншими словами, перехресна перевірка дає хороші оцінки точності передбачення моделі.

Кожен з алгоритмів класифікації має свої переваги та недоліки, а мета полягає в тому, щоб створити межі рішень. Наприклад, дерево рішено було обрано на етапі побудови моделі для запропонованої системи, оскільки він забезпечує наступні переваги: простота, інтерпретація та ефективність. Відповідні ознаки бажаного місця призначення кожного туриста (наприклад, національність, дохід сім'ї тощо) використовуються для побудови моделі, яка описує переваги користувача. Метод опорних векторів – алгоритм класифікації, який має суттєве теоретичне доведення та успішно застосовується в багатьох реальних програмах, (розпізнавання обличчя, розпізнавання тексту тощо). Метод опорних векторів – це контрольований алгоритм машинного навчання, який спочатку був розроблений для використання в двійковій класифікації. Концепція методу полягає в ідеї пошуку оптимальної гіперплощини, яка може розрізнити набір даних на два класи. Метод опорних векторів – це ще один контрольований алгоритм машинного навчання, який розширює концепцію єдиного персептрона. Метод опорних векторів, працює з мережею, яка складається з одного вхідного рівня, одного вихідного рівня та довільної кількість прихованих шарів, розташованих між вхідним і вихідним шарами. Дані переміщуються від вхідного рівня через приховані вузли до вихідних вузлів. У мережі використовується функція активації або передачі – це функція, яка перетворює набір вхідних сигналів у вихідний сигнал.

Порівняти продуктивність нашої системи з іншими існуючими системами складно з кількох причин, зокрема через кількість пунктів призначення, різні міста та місця, критерії ефективності та відмінності в методах оцінки.

Детально дослідимо вказані причини:

1. Кількість користувацьких введень і кількість пунктів призначення

Існуючі рекомендаційні системи спрямовані на підвищення точності системи та ігнорують практичні аспекти. Наявність високої точності рекомендацій, зумовлена необхідністю великої кількості вхідних даних від користувача, не обов'язково означає, що система рекомендацій достатньо розвинена.

2. Місто та розташування

Місто або місце розташування, яке застосовує система рекомендацій, відіграє важливу роль у її ефективності. Кожне місто має свою унікальну і складну природу. Використання тих самих факторів для пов'язування з різними пунктами призначення може призвести до різних результатів. Наприклад, поведінка туристичних витрат може не корелюватися з процесом пошуку місць призначення в деяких країнах, але цей фактор може виявити високу кореляцію в інших країнах.

3. Методи оцінювання

Більшість існуючих рекомендаційних систем не надають жодних методів перевірки своїх систем. Найкращий спосіб оцінити рекомендацію – це використовувати методи тестування.

Одним із багатообіцяючих способів вирішення складних проблем у реальному житті є вибірка голосів кількох експертів з подальшим остаточним рішенням, отриманим шляхом об'єднання їхніх голосів. Ця концепція також застосовується в машинному навчанні та відома як ансамбль класифікаторів або ансамблеве навчання. Цей метод є алгоритмом навчання з учителем, який використовує комбіновані моделі для отримання вищої точності класифікації. Було показано, що ансамблеве навчання потенційно покращує продуктивність і надійність прогнозування, але це не гарантовано.

Отже в роботі пропонується гібридний підхід на основі ансамблю для підвищення ефективності системи на основі моделі з точки зору ефективності

класифікації. Результати класифікації, такі як прогнозування, оцінка ймовірності та ранжирування з алгоритмів класифікації, об’єднуються для отримання єдиного та більш надійного кінцевого результату.

Метою проведеного дослідження є покращення ефективності класифікації запропонованої системи шляхом вивчення інших традиційних алгоритмів класифікації, включаючи дерево рішень, метод опорних векторів і багатошаровий перцептрон. Ефективність класифікаторів оцінюється за допомогою восьми наборів даних про вибір місця призначення, які були створені при проведенні дослідження.

Усі змінні в наборах даних є категоріальними змінними (наприклад, порядкові, номінальні), і було помічено, що ці типи змінних можуть спричиняти розривний зв’язок між незалежною змінною та залежною змінною. Щоб підготувати дані для класифікаторів, номінальні та порядкові змінні як для входів, так і для виходів необхідно перетворити на числові змінні, інакше вони можуть призвести до неправильного модель.

У цьому дослідженні використовуються три класифікатори: дерево рішень, метод опорних векторів та багатошаровий персептрон. Було проведено дослідження ефективності класифікації методом опорних векторів та багатошарового персептрона. Розглянемо детальніше запропоновані методи машинного навчання.

Дерева рішень – це універсальні алгоритми машинного навчання, які можуть виконувати завдання класифікації і регресії з кількома виходами [1]. Це потужні алгоритми, які здатні обробляти складні набори даних.

Дерева рішень є також фундаментальними компонентами випадкових лісів, одних із найпотужніших алгоритмів машинного навчання. Однією з ключових особливостей дерев рішень є те, що вони не вимагають підготовки даних. Фактично вони взагалі не вимагають масштабування або центрування елементів. Дерева рішень також можуть оцінювати можливість того, що екземпляр належить певному класу *k*. Спочатку алгоритм переглядає дерево, щоб знайти ключовий вузол для

цього екземпляра, а потім повертає співвідношення навчальних екземплярів класу *k* у цьому вузлі.

Дерева рішень мають багато переваг:

1. Вони прості для розуміння та інтерпретації.
2. Прості у використанні.
3. Універсальні.
4. Потужні.

Проте вони також мають кілька обмежень:

1. Деревам рішень підходять ортогональні межі рішень, які роблять їх чутливими до ротації навчальних наборів.
2. Дерева рішень схильні до перенавчання під час роботи із завданнями регресії.

Метод опорних векторів

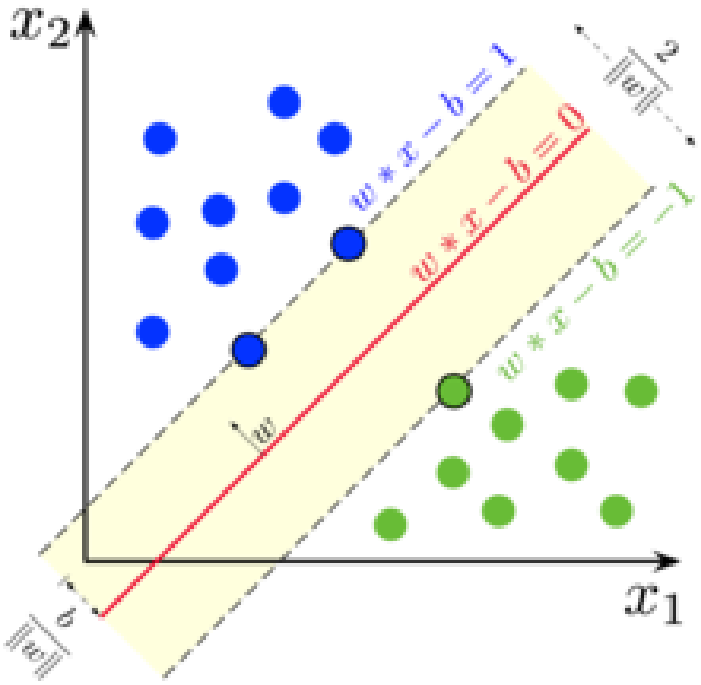


Рисунок 2.1 – Максимально розділова гіперплощина та межі [24]

Метод опорних векторів зазвичай використовується для вирішення проблем класифікації та регресії. Цей метод успішно застосовувався в багатьох областях для вирішення завдань класифікації, таких як розпізнавання цифрових символів

рукописного тексту, виявлення обличчя тощо. Цей підхід проектує вхідні дані у простори з більшою вимірністю, щоб можна було розділити нелінійні дані. Метою є оптимізація гіперплощини, яка полягає в максимізації відстані від кожної точки до гіперплощини, як показано на рис. 2.1. Метод опорних векторів складається з двох основних фаз. По-перше, функція ядра використовується для відображення даних у вищий вимір (тобто лінійна, поліноміальна, функція зміщення радіуса). У цей момент гіперплощина може бути використана для розділення двох класів. Для набору даних, який неможливо ідеально розділити лінійно, мета процесу полягає в тому, щоб знайти набір ваг, які визначають дві гіперплощини, як визначено нижче:

$$\begin{aligned} \vec{w}\cdot\vec{x}+b &\geq +1 \\ \vec{w}\cdot\vec{x}+b &\leq -1 \end{aligned} \tag{2.7}$$

У випадку нелінійно роздільних даних метод опорних векторів може обробляти нероздільні точки, вводячи слабкі змінні, як показано нижче:

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \tag{2.8}$$

У цьому дослідженні гауссове ядро радіальної базисної функції було обрано як найбільш прийнятну функцію ядра, оскільки наш набір даних складається з невеликої кількості ознак, а радіальна базисна функція використовує менше гіперпараметрів, ніж поліноміальне ядро. Для цього дослідження було обрано RBF Гауса, як визначено в рівнянні:

$$f\left(x_i\right) = \exp\left(-\frac{1}{(2\sigma^2)}\left\|x_i - x_j\right\|^2\right) \tag{2.9}$$

Щодо переваг методу опорних векторів, то цей класифікатор здатний знаходити глобальний мінімум, а його проста геометрична інтерпретація створює необхідну базу для майбутніх досліджень. Найбільш вигідною характеристикою

нелінійного класифікатора методу опорних векторів є опуклість. Однак метод опорних векторів також має кілька недоліків: він дуже чутливий до параметрів ядра та вибору ядра; отже, вибір параметра, який трохи виходить за межі, може призвести до низької ефективності класифікації.

Налаштування цих параметрів зазвичай необхідне для хорошої продуктивності. Наприклад, вибір параметра вартості є критичним. Використання більшої вартості може призвести до перепідгонки моделі. Крім того, розробка моделі за допомогою методу опорних векторів вимагає трудомісткого підходу проб і помилок і займає досить багато часу, особливо для великого обсягу даних. Нижче представлено алгоритм, який використовувався для оптимізації гіперпараметрів у методі опорних векторів у цьому дослідженні:

1: Input: trainD, $\log_2$ c_vector, $\log_2$ g_vector

2: Output: *w**(c,γ) % Large scale search

3: stepsize = 1;

4: for i=1:numl($\log_2$ c_vector) % loop through every element in the list.

5: for j=1:numel($\log_2$ g_vector)

6:),,,(maxarg),(*trainwDcwCVcwλγ=

7: if w*>best w

8: c*=c,g*=g;

9: end

10: end

11: end

12: stepsize = prev_stepsize ÷ 2;% Adjust the medium-scale and small-scale search

13: $\log_2$ c_vector = c*-prev_stepSize:stepsize:g*+prev_stepsize;

Процедура побудови моделі методу опорних векторів виглядає наступним чином:

1. Провести масштабування на наборах даних.
2. Перебір різних ядер (лінійне, поліноміальне).
3. Виконати перехресну перевірку сітки пошуку, щоб знайти оптимальні параметри.

4. Навчання моделі за допомогою навчального набору з отриманими оптимальними параметрами.

5. Перевірка за допомогою тестового набору даних і оцінка за допомогою показників ефективності.

Багатошаровий персептрон

Багатошаровий перцептрон вважається прямою мережею, універсальним апроксиматором. Це найбільш поширений метод у галузі штучних нейронних мереж для вирішення завдань класифікації. Нейронну мережу можна навчити передбачати змінну класу. Існує багато типів штучних нейронних мереж, які використовуються для класифікації, включаючи багатошаровий перцептрон, радіальну базову функцію та ймовірнісні нейронні мережі. У цьому дослідженні як тип мережі було обрано саме багатошаровий перцептрон; її архітектура складається з одного або кількох прихованих шарів між вхідними та вихідними вузлами, і кожен із вузлів у мережі підключений і має певну вагу. Рис. 3 ілюструє загальну мережеву архітектуру перцептрону. Багатошаровий перцептрон відображає дані з простору ознак у вихідний простір класифікації, і прогноз можна вибрати як вектор кодування, який є найближчим до виходу (тобто вихід, який відображає найвище значення, є класом-переможцем).

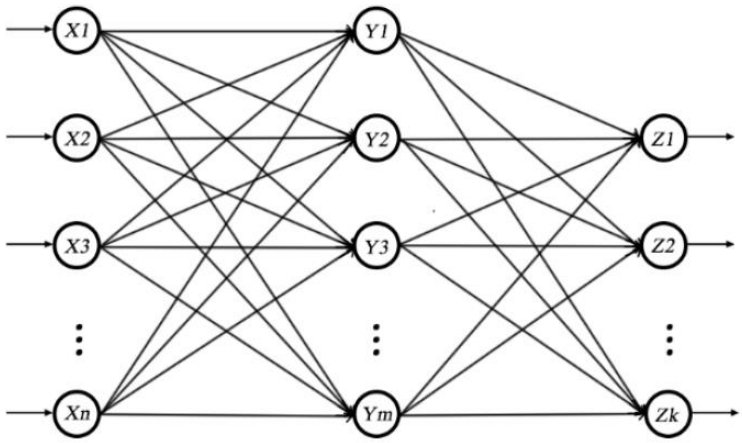


Рисунок 2.2 – Архітектура багатошарового персептрона з одним прихованим шаром

Модель багатошарового персептрону була розроблена з використанням наступних критеріїв:

1. Архітектура мережі складається з одного вхідного рівня, одного прихованого рівня та одного вихідного рівня. Вхідний рівень містить вхідні вектори, і тут не виконуються обчислення. У прихованому шарі вибирається один прихований шар. Більшість складних проблем можна вирішити за допомогою одного прихованого шару. Вихідний рівень містить вихідний вектор, до якого застосована функція активації.

2. Вибір кількості прихованих вузлів. Наскільки нам відомо, не було зроблено висновку щодо кількості прихованих нейронів, які слід використовувати в прихованому шарі; отже, кількість оптимальних прихованих вузлів базується на процесі проб і помилок. Рішення щодо кількості прихованих нейронів у прихованому шарі має вирішальне значення, оскільки це може призвести до надмірної підгонки та довшого часу обчислень, якщо використовуємо занадто багато прихованих нейронів або недостатню підгонку, коли нейронів у прихованому шарі занадто мало.

3. Функція softmax була використана як функція активації для всіх шарів і задач бінарної та багатокласової класифікації. Функція гарантує, що сума ймовірностей усіх класів дорівнює 1. Враховуючи, що ми маємо вектор x із K результатів, функцію можна обчислити як:

$$f(x_i) = \frac{e^x}{\sum_{j=0}^{K} e^x}, i = 0..K \tag{2.10}$$

Алгоритм зворотного розповсюдження Scaled Conjugate Gradient (SCG) використовувався в цьому дослідженні під час навчання мережі. Вважається, що SCG є кращим за стандартний алгоритм зворотного поширення, оскільки він усуває

деякі важливі недоліки, такі як низька швидкість збіжності та залежність від параметрів користувача. Мережа була навчена та перевірена 10 разів через недоліки штучної нейронної мережі, які страждають від кількох локальних мінімумів. Вибрано мережу, яка показала найвищий рівень точності. Коротко кажучи, процедура відбору та оцінки моделі була наступною:

1. Провести масштабування входу та виходу.

2. Використати перехресний пошук оптимальної кількості прихованих нейронів із співвідношення 1:1:100.

3. Використати перехресний пошук оптимальної кількості прихованих нейронів, використовуючи емпіричне правило.

4. Навчання мережі з отриманою оптимальною кількістю прихованих нейронів.

5. Тестування з тестовими даними та оцінювання за допомогою показників ефективності.

3 ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ЗАПРОПОНОВАНОЇ СИСТЕМИ ТА РОЗРОБКА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ДЛЯ РЕКОМЕНДАЦІЇ ТУРИСТИЧНИХ МАРШРУТІВ

3.1 Представлення набору даних

На рис. 3.1 представлено розподіл класів для 20 напрямків з території України. З графіка видно, що це незбалансований набір даних, оскільки розподіл класів є нерівномірний. Одним із завдань у цьому дослідженні була розробка моделі, яка б працювала з такими даними. Модель, яка була побудована з використанням усіх 20 пунктів призначення, досягла дуже низького рівня точності класифікації в 17%, була складною і потребувала багато часу для створення.

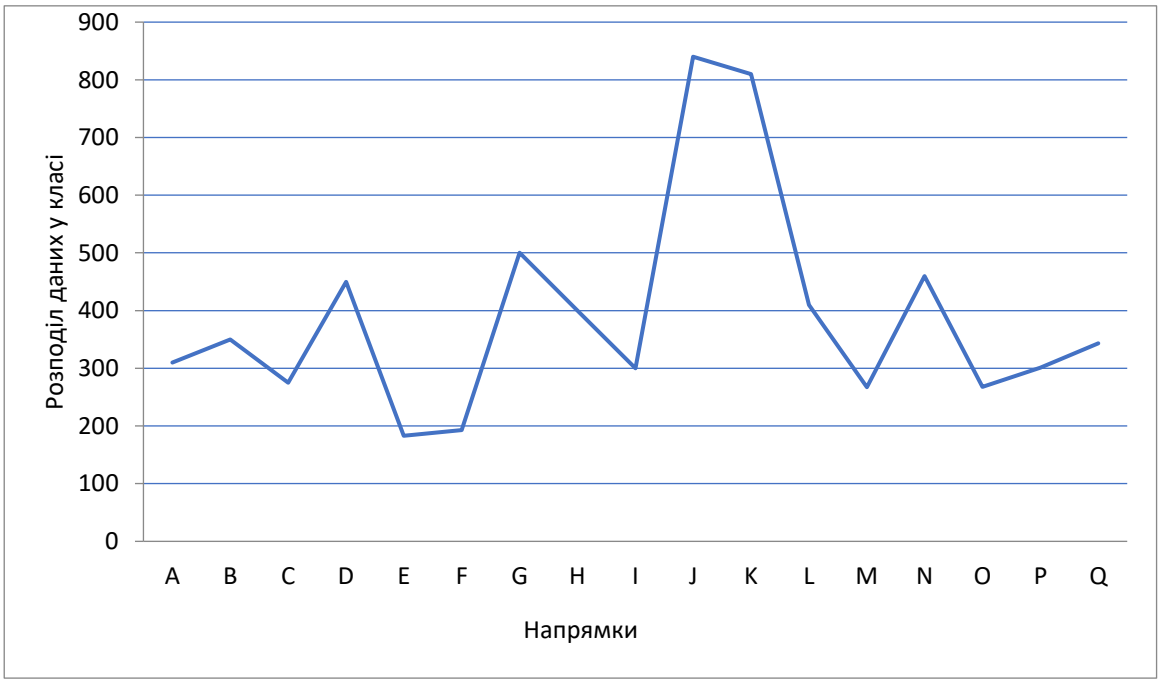


Рисунок 3.1 – Розподіл класів для набору даних

Модель була надто складною, оскільки мала великий розмір дерева. Це ускладнювало інтерпретацію для особи, яка приймає рішення. Щоб вирішити цю проблему, в досліджені було запропоновано застосовувати декомпозицію класу на етапі попередньої обробки. Мета полягала в тому, щоб ідентифікувати групи

напрямків зі спорідненими моделями. Декомпозиція класів дає нам багато переваг, включаючи підвищену продуктивність класифікації, масштабованість до великої бази даних, підвищену зрозумілість, модульність і придатність для паралельних обчислень.

Вибір оптимального методу декомпозиції для певного типу задачі класифікації складний. Існує багато методів декомпозиції класів, таких як кластеризація з *k*-середнім, кодова матриця, агрегація концепцій тощо. У зв'язку з тим, що в роботі враховувалася взаємодія з користувачем і значення нової кластерної групи/категорії призначення, проблема класифікації 20 мультикласів була чітко розкладена на кілька підпроблеми, досліджуючи типи бажаних місць призначення туристів (поєднуючи знання експертів у сфері туризму та інформацію про місце призначення з веб-сайту). Було створено десять категорій призначення та застосовано розподіл за класами. Моделі були побудовані на основі категорій призначення, які представлені в більш ніж одному класі (тобто набір даних, який представляє проблему бінарної або багатокласової класифікації). Що стосується характеристик кожного набору даних, категорія «Природа» складається з трьох класів (два з них представляють водоспади, а один — озеро); а категорія «Музей і художня галерея» складається з двох класів (оскільки є як спеціалізовані музеї, так і художні галереї).

Таким чином, було створено десять категорій туристичних напрямків. Моделі були налаштовані на основі категорій, які мали більше ніж один клас.

Після початкового вибору вхідних даних неперервні змінні були дискретизовані за допомогою методу групування. Викиди були виявлені за допомогою запропонованого нижче простого алгоритму. Порядкові змінні були зменшені від 5 до 3. Деякі змінні були нормалізовані з використанням експертних знань у галузі туризму

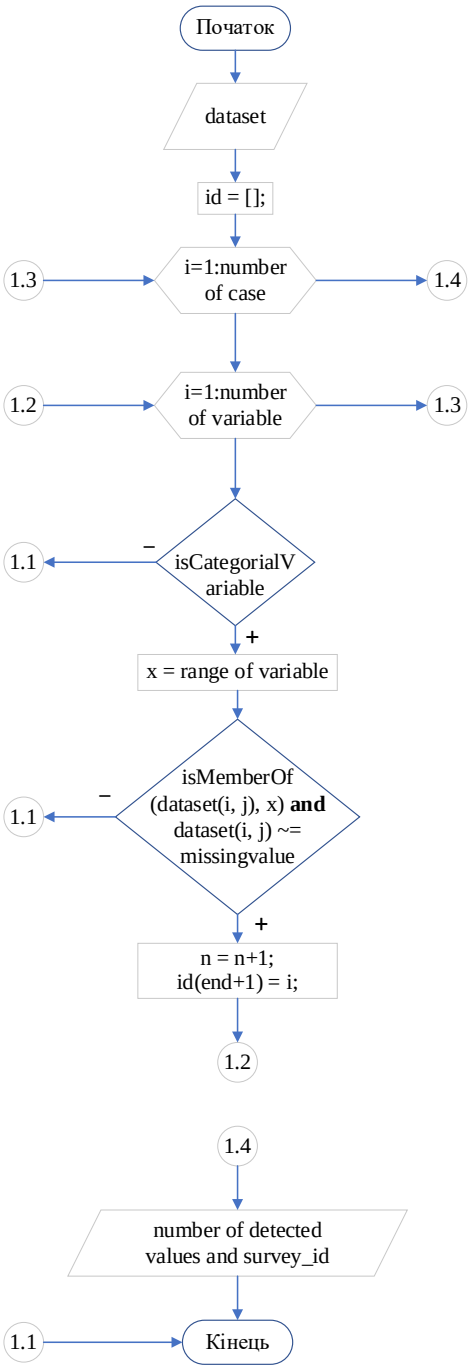


Рисунок3.2 – Алгоритм виявлення викидів

Після того, як набір даних було очищено та перетворено, був застосований метод двоетапної фільтрації до процесу редукції даних. Це було зроблено, щоб видалити нерелевантні та зайві функції з набору даних.

Наступним кроком необхідно виміряти подібність в процесі вибору ознак, щоб охарактеризувати як релевантність, так і надмірність змінних. У рівнянні (3.1) нам задано набір X і Y, $p(x)$ та $p(y)$ граничні функції розподілу ймовірностей змінних X і Y, $p(x,y)$ спільна функція розподілу ймовірностей X і Y:

$$MI(X,Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.1}$$

Однак, використовуючи безперервні змінні, спільну ймовірність і граничну ймовірність важко оцінити. На практиці неперервні змінні часто дискретизують на дискретні змінні, а потім необхідну інформацію можна обчислити за допомогою наступного рівняння:

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.2}$$

де *p(x,y)*– спільна ймовірність, яка є ймовірністю того, що дві змінні відбудуться одночасно, *p(x)* чи *p(y)* – гранична ймовірність або ймовірність появи однієї змінної.

Наступним кроком є виконання першої фільтрація. Метою першого кроку фільтрації є ранжування змінних і видалення будь-яких незалежних змінних, які не пов’язані із залежною змінною. Був застосований алгоритм вибору ознак максимальної релевантності представлений нижче, у якому обрана взаємна інформація як вимірювання для видалення нерелевантних ознак. Обчислили оцінку взаємної інформації між кожною незалежною та залежною змінною. Потім ранжували їх у порядку спадання та використали порогове значення (вибране вручну), щоб видалити функції, які мали менший внесок або не були пов’язані з прогнозною потужністю:

$$max D(S,c), D = MI(\{x_1, i = 1,...,t\}; c) \tag{3.3}$$

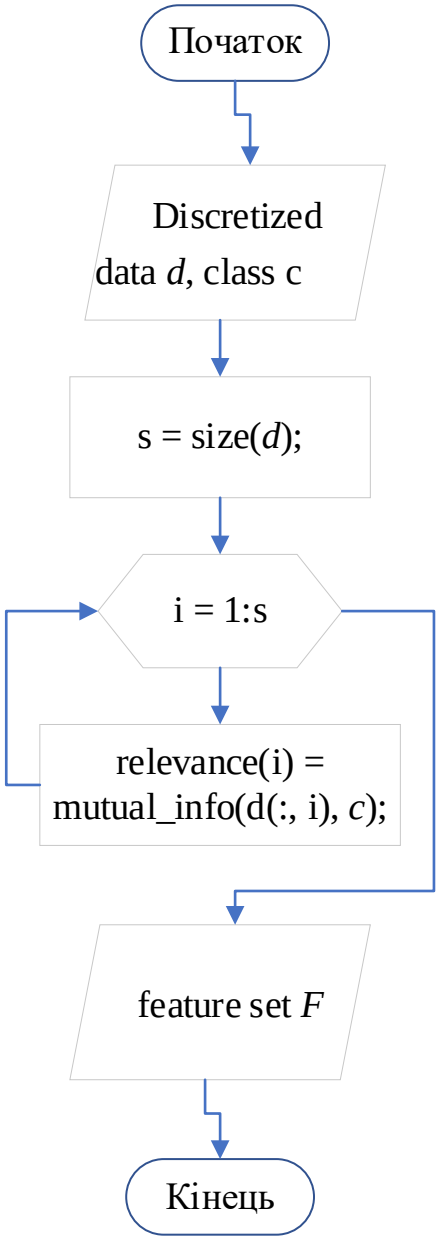


Рисунок 3.3 – Алгоритм максимальної релевантності

На етапі вибору функції було виконано перший метод фільтрації. Використовувалась різна кількість порогових значень на основі кожного набору даних, щоб вибрати 10% ознак.

Рисунок 3.4 – Значення взаємної інформації для кожної категорії

На другому етапі фільтрації використовувалися два алгоритми вибору ознак на основі взаємної інформації: мінімальна надлишковість і максимальна релевантність і нормалізований взаємний інформаційний вибір ознак, щоб видалити зайві змінні з набору даних.

Алгоритм мінімальної надлишковості та максимальної релевантності використовує значення взаємної інформації для ранжування ознак на основі мінімальної надмірності та максимально відповідних критеріїв. Алгоритм обчислює надмірність для кожної пари функцій і релевантність між функціями та класом. У цьому дослідженні було розглянуто алгоритми взаємної інформації для дискретних змінних і у формі:

$$MRMR = \max_{1 \in \Omega_S} \left[I(i,h) - \frac{1}{|S|} \sum_{j \in s} MI(i,j) \right] \tag{3.4}$$

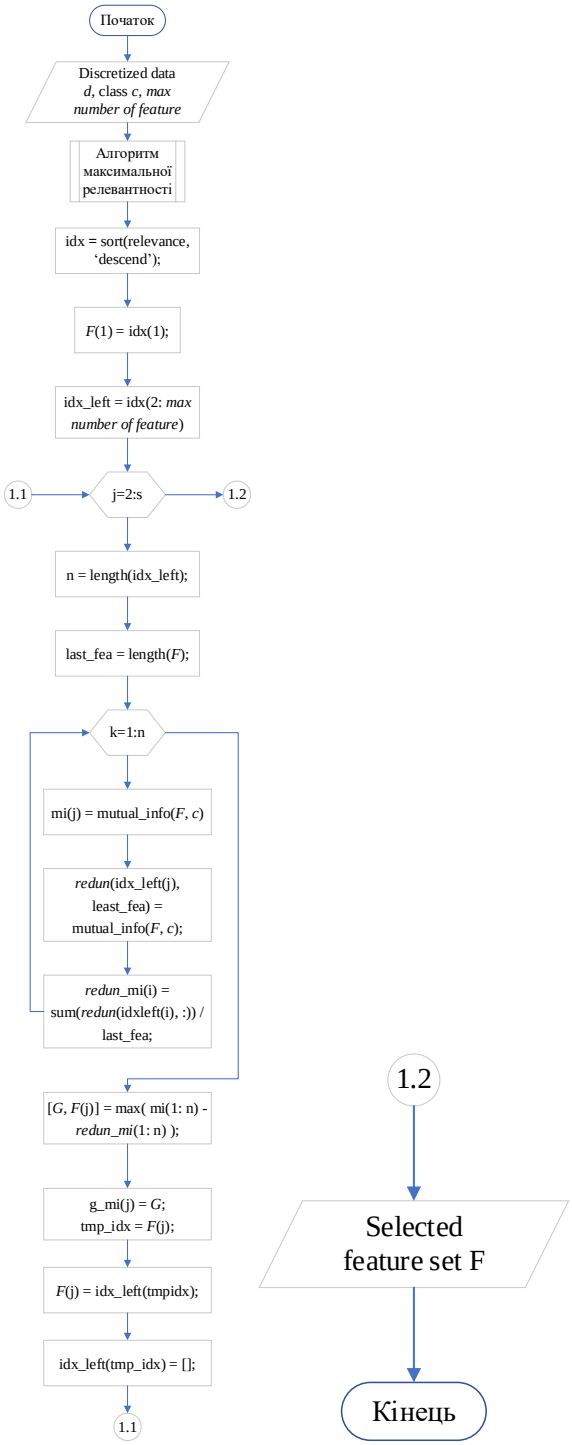


Рисунок 3.5 – Алгоритм мінімальної надлишковості та максимальної релевантності

Алгоритм мінімальної надлишковості та максимальної релевантності можна дещо модифікувати, що покращить його роботу. Модифікований алгоритм нормалізує початкове значення взаємної інформації за допомогою мінімальної ентропії (*H(i)* і *H(j)*) обох ознак, як показано в рівняннях (3.5) і (3.6).

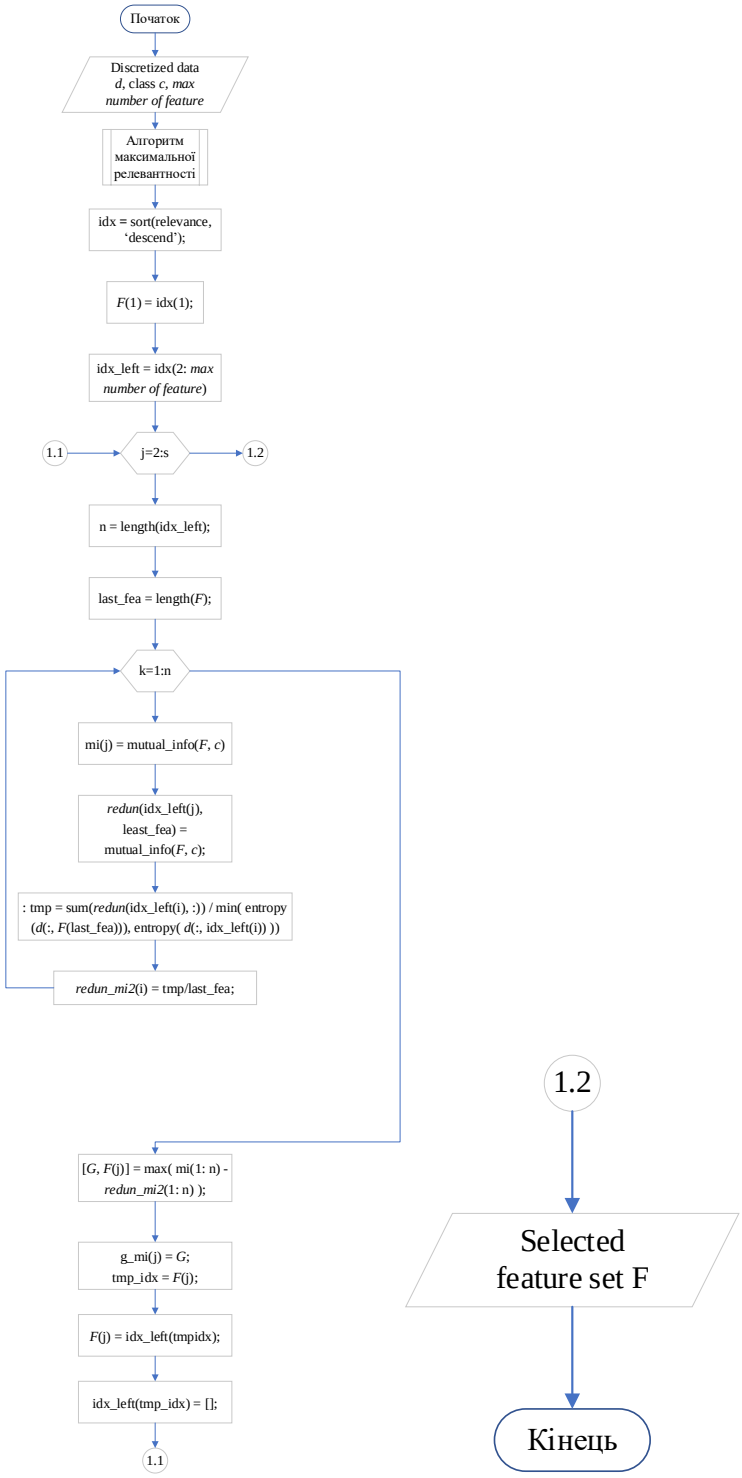


Рисунок 3.6 – Модифікований алгоритм нормалізованого вибору взаємної інформації

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{3.5}$$

Тоді модифікацію взаємної інформації для отриманого алгоритму можна записати наступним чином:

$$MI2(i,j) = \frac{MI(i,j)}{\min\{H(i), H(j)\}} \tag{3.6}$$

Отже, остаточний вигляд модифікованого алгоритму можна записати у вигляді рівняння нижче:

$$NMSFS = \max_{1 \in \Omega s} \left[I(i,h) - \frac{1}{|S|} \sum_{j \in s} MI2(i,j) \right] \tag{3.7}$$

Модифікований алгоритм нормалізованого вибору взаємної інформації представлений на рис. 3.6.

Аспекти алгоритму класифікації:

1. Підхід до класифікації на основі правил.

Багатообіцяючим підходом до підвищення точності класифікації є використання класифікаторів на основі правил, тому що це дає можливість отримати користь від правил, отриманих з моделей. Правила можна скоротити за допомогою експерта з туристичної сфери для отримання більшої точності прогнозування. Крім того, нерелевантні або зайві функції також можна усунути під час процесу перетворення з дерева рішень на правила шляхом інтеграції існуючого алгоритму або модифікації алгоритму.

2. Підхід до глибокого навчання.

Ще одна нова парадигма в машинному навчанні – глибоке навчання. Глибоке навчання було успішно застосовано в програмах комп’ютерного зору, таких як розпізнавання зображень. Було б цікаво побачити, як цей метод машинного навчання можна використовувати в категоріальних наборах даних, подібних до нашого, які етапи попередньої обробки даних будуть необхідні перед навчанням

моделі та яким буде вибір архітектури мережі для проблеми класифікації призначення.

3.Аспект інтерфейсу користувача.

Існує три напрямки дослідження запропонованого інтерфейсу користувача:

1. До семантичних веб-сайтів.

Перший передбачає подолання розриву між згенерованим файлом моделі, таким як XML і JSON, і мовою семантичних веб-правил.

2. Механізм зворотного зв'язку.

Другий напрямок – впровадження механізму зворотного зв’язку, щоб турист міг оцінити та залишити відгук про напрямки. Інтеграція відгуків і оцінок користувачів може покращити нашу рекомендаційну систему. Відгуки та оцінки користувачів можна використати за допомогою аналізу тексту для створення ефективнішого інтерфейсу користувача.

3. Оцінка інтерфейсу користувача.

Подальший розвиток користувальницького інтерфейсу для запропонованої рекомендаційної системи має зосередитися на методі оцінки. Методи, що включають евристичне оцінювання, тестування зручності використання, рекомендації та когнітивне керівництво, слід ретельно переглянути, оскільки кожен метод оцінювання має свої переваги та недоліки. Метод рекомендацій вважається найкращим для пошуку загальних і повторюваних проблем. Однак цей метод має проблеми при виявленні серйозних проблем.

У цій магістерській роботі запропонована інтелектуальна рекомендаційна система, яка використовує підходи на основі моделі та ансамблю, що грунтуються на методах машинного навчання. Було проведено порівняння кількох добре відомих алгоритмів класифікації та виявлено, що багатошаровий персептрон перевершує інші щодо наборів даних. На експериментальному дослідженні представлено, як можна використати методи ансамблевого навчання для підвищення рівня точності класифікації рекомендаційної системи. Крім того, розроблена інформаційна технологія заснована на моделі інтерфейсу користувача,

який має адаптивні, чуйні та інтерактивні можливості, була проведена в кінці цієї дипломної роботи, щоб підвищити рівень задоволеності користувача системою.

У цій роботі також досліджувалися п’ять наборів факторів, які вплинули на обрані туристами місця призначення, включаючи характеристики поїздки, характеристики туриста, поведінку туристів щодо витрат, мотивацію до подорожі та соціально-демографічну інформацію туристів на основі якісних досліджень. Поведінка туристичних витрат є найважливішим фактором при класифікації. Тридцять п'ять ознак були виявлені як такі, що мають найбільший вплив на запропоновану рекомендаційну систему.

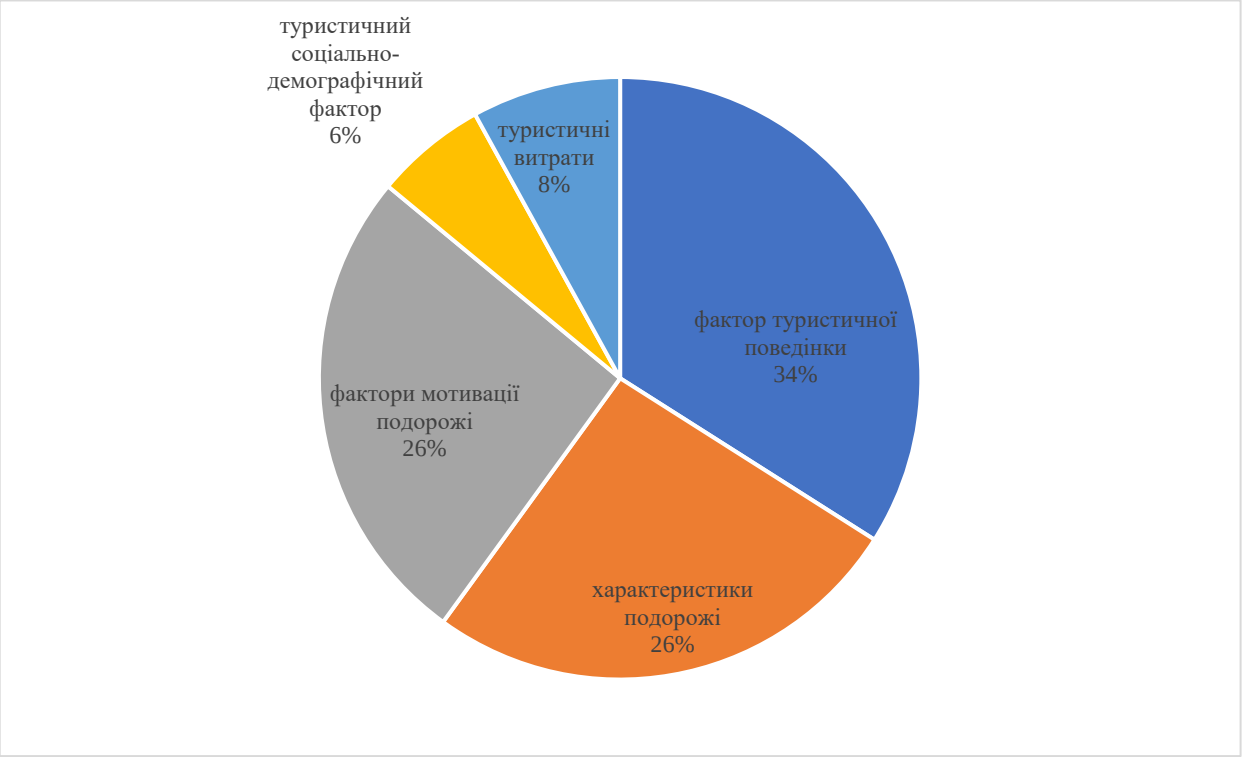


Рисунок 3.7 – Розподіл головних факторів, які були використані в моделях вибору місця призначення

Рис. 3.7 ілюструє внесок факторів, що є важливими у моделях вибору пункту призначення. Можна побачити, що фактор туристичної поведінки був найбільш часто використовуваним (34%), за яким йшли характеристики подорожі (26%) і фактори мотивації подорожі (26%). Туристичний соціально-демографічний фактор робить найменш значний внесок у систему (6%). Експериментальні результати

також підтверджують висновки з літератури, які вказують, що поєднання факторів туристичної мотивації допомагає підвищити точність класифікації, оскільки цей фактор був визначений як такий, що має найбільший вплив і використовувався в моделі як найбільш актуальна функція.

З точки зору практичних аспектів, запропонована рекомендаційна система використовувала невелику кількість відповідних і ненадлишкових вхідних даних із 3–5 функцій для досягнення найкращих результатів рекомендацій. Це означає, що запропонована система вважається ненав'язливою та, ймовірно, буде прийнята користувачами. Створені моделі можуть допомогти особам, які приймають рішення, отримати огляд кількох етапів, які слідуватимуть за кожним можливим рішенням під час вибору пункту призначення в Чіангмаї. Крім того, правила прийняття рішень з оптимальних моделей були витягнуті для того, щоб особам, які приймають рішення, було легше зрозуміти результати, які показують, що Temple-landmark і Temple-peaceful мали найменше правил. Ці правила використовуватимуться під час інтеграції онлайн-фази в систему.

3.2 Методика рекомендації туристичних маршрутів

Запропонована система базується на трирівневій архітектурі веб-моделі, більш відомої як клієнт-серверна архітектура. Архітектура, яка складається з трьох рівнів, складається з рівня презентації, додатків і даних. Презентаційний рівень – це інтерфейс користувача, реалізований за допомогою технології веб-браузера, за допомогою якого він отримує вхідні дані, такі як демографічні дані, характеристики користувачів і вимоги користувачів від туристів, і відображає результати користувачам. Другий шар – це прикладний рівень, який виконує роль середнього шару. Він відповідає за оптимізацію та логічне прийняття рішень, а також за оцінку даних та інші розрахунки. Рівень даних приймає та зберігає всю інформацію з верхніх рівнів. Інформація та відповідні дані, як-от географічні дані та інформація про подорожі користувача, зберігаються на різних рівнях за допомогою форматів файлів eXtensible Markup Language (XML) і JavaScript Object

Notation (JSON). У цьому дослідженні запропоновано інтерфейс користувача для запропонованої рекомендаційної системи, який має адаптивні, чутливі та інтерактивні можливості.

Адаптація для інтерфейсу користувача повинна включати деякі фактори, такі як продуктивність користувача, цілі користувача, когнітивне навантаження, обізнаність користувача про ситуацію, знання користувача, профілі груп, змінні ситуації та змінні завдання. Дерево рішень можна використовувати як алгоритм адаптації та як один із методів адаптації інтерфейсу. В інтерфейсі користувача швидкість реагування стосується змін розміру вікна браузера та способу розташування вмісту.

Інтерактивність є одним із найбільш багатообіцяючих аспектів, які слід розглянути, щоб використати весь потенціал рекомендаційної системи. Розробка та впровадження справжнього інтерактивного веб-сайту потребує великої роботи, яка передбачає спільне ставлення користувачів, чіткий процес і стандарти для управління вмістом, а також дослідження дизайну. У цьому дослідженні було поставлено завдання збільшити інтерактивність між користувачем і системою, щоб відображати корисну інформацію (наприклад, місце призначення) користувачам за допомогою інтерактивних карт. Крім того, передові веб-технології, такі як JQuery, CSS і HTML5, можуть бути використані для покращення взаємодії з користувачем і підвищення відгуку та інтерактивності системи.

Запропонована рекомендаційна система призначена для використання туристами та турагентами та складається з онлайн- та офлайн-фаз. У офлайн-фазі система виконує розрахунок моделей оптимального вибору напрямків, щоб рекомендувати напрямки туристам, заощаджуючи туристам додаткові витрати на використання обладнання та час під час процесу пошуку інформації. Необроблені дані, наприклад записи обстежень, подаються в систему через модуль керування даними. Цей модуль відповідає за інтеграцію, очищення, перетворення, зберігання та підтримку даних опитування. Обслуговування системи просто вимагає введення нових даних у систему рекомендацій у модулі керування даними на цьому рівні. Наприклад, щороку, коли отримуються нові дані опитування, їх можна інтегрувати

у існуючий набір даних, і відповідно створюватимуться нові моделі, які передаватимуться на веб-сервер у верхньому рівні. У модулі керування інтерфейсом користувача можна додавати, редагувати, видаляти або змінювати моделі.

У модулі «Керування моделлю» встановлено класифікатори дерево рішень та інші класифікатори машинного навчання, включаючи три добре відомі алгоритми класифікації: дерево рішень, метод опорних векторів, багатошаровий персептрон та інші моделі ансамблевого навчання. Вони використовуються для розрізнення конкретних пунктів призначення в кожному наборі даних. Щоб зробити комплексну модель придатною для використання та інтерпретувати її результати для туриста, моделі дерева рішень перетворюються на правила прийняття рішень, а потім інформація передається до модуля керування інтерфейсом користувача (рис. 3.8.).

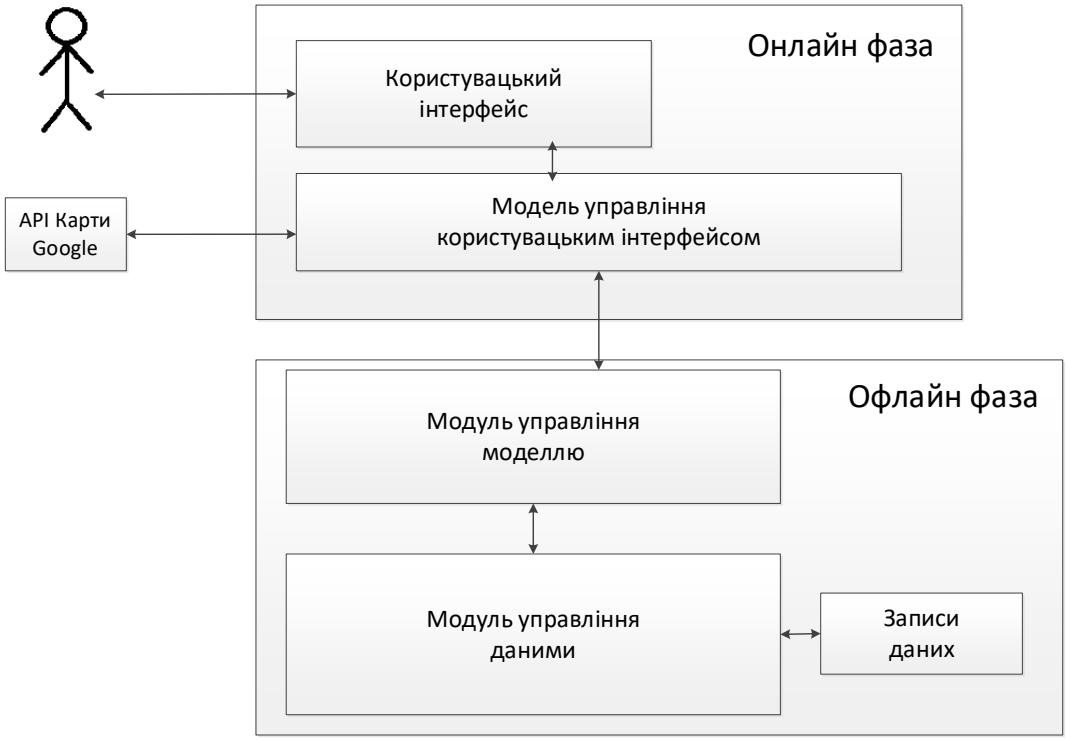


Рисунок 3.8 – Практична система рекомендацій для туристів

На онлайн-фазі верхній рівень можна вважати клієнтським, оскільки він містить інтерфейс користувача, де турист може взаємодіяти з системою через різні

платформи, такі як мобільний, робочий стіл або веб-браузер. У модулі керування інтерфейсом користувача правила прийняття рішень перетворюються у формати XML і JSON для створення нового інтерфейсу користувача. Крім того, система може підключатися до Google API для отримання відповідної інформації, яка стосується карт і маршрутів, щоб система могла відображати результати в інтерфейсі. Туристи можуть взаємодіяти з системою через інтерфейс користувача. Щоб отримати рекомендований пункт призначення, турист повинен надати ряд вхідних даних, наприклад стиль поїздки та межі вартостей, а також інші в систему, вибравши відповіді зі списків. Згодом рекомендовані результати включатимуть назву пункту призначення та маршрут подорожі, який буде отримано за допомогою наданої інформації, отриманої з місця розташування користувача та вибраного пункту призначення. У цьому шарі зберігається географічна, просторова та маршрутна інформація. Система підключається до кількох API Google, таких як GMap і GLargeMap, щоб мати можливість завантажувати карти та керувати ними.

3.3 Інформаційна технологія рекомендації туристичних маршрутів

У пошуках дослідження та аналізу результатів різних етапів запропонованої рекомендаційної системи було досліджено два існуючих прототипи рекомендаційних систем, а саме: персоналізовану систему планування подорожей та інтелектуальну систему туристичних атракцій. Цілі цього техніко-економічного обґрунтування пояснюються нижче.

Першою метою цього техніко-економічного обґрунтування було виявити існуючі проблеми в розробці рекомендаційних систем через розроблений прототип і експерименти, а також визначити, чи можливо замінити баєсівську модель запропонованою в роботі моделлю дерева рішень в системі рекомендацій. Друга мета полягала в тому, щоб порівняти існуючі вимірювання подібності з попередніми рекомендаційними системами, які мали подібні типи набору даних, і визначити, чи можливо використовувати взаємну інформацію як вимірювання подібності.

Після проведення відповідних досліджень та обчислень було розроблено інформаційну технологію рекомендації туристичних маршрутів, яка має наступну архітектуру (рис.3.9).

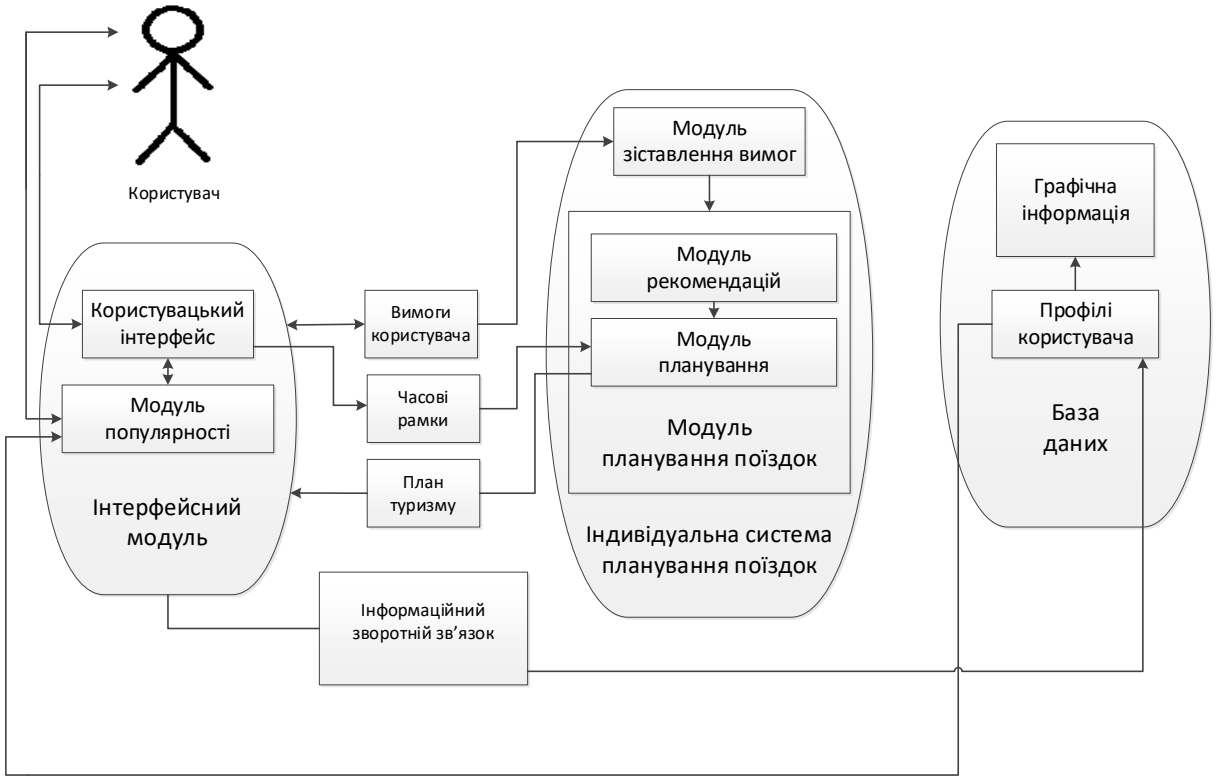


Рисунок 3.9 – Загальна структура інформаційної технології

Як показано на рис. 3.9, модулі, на яких було зосереджено це дослідження, це модуль бази даних і модуль індивідуальної системи планування подорожей, а також алгоритм обґрунтування розкладу, який використовується для створення персоналізованого розкладу подорожей із кінцевого набору туристичних послуг, що включають місця визначних пам'яток, заклади харчування та ресторани, варіанти розміщення, розташування готелів, вимоги користувачів тощо. Алгоритм включає кілька кроків для пошуку місця подорожі чи пункту призначення та розрахунки, пов'язані з транспортуванням і часом перебування. Механізм зворотного зв'язку – це метод, який використовується для ранжування цікавих місць (готелі, ресторани, житло), який є сукупним значенням рейтингів популярності користувачів.

Модуль зіставлення вимог щодо подорожі відповідає введеним користувачам (наприклад, необхідним пам’яткам, готелям, ресторанам із бази даних). Потім рекомендований модуль виконав дії:

1. Пошук місця подорожі або пункту призначення.
2. Розрахунок транспортування та часу проживання.
3. Додавання обраного місця подорожі в часові рамки.

Інтерфейс рекомендаційної системи представлений на рис. 3.10-3.13.

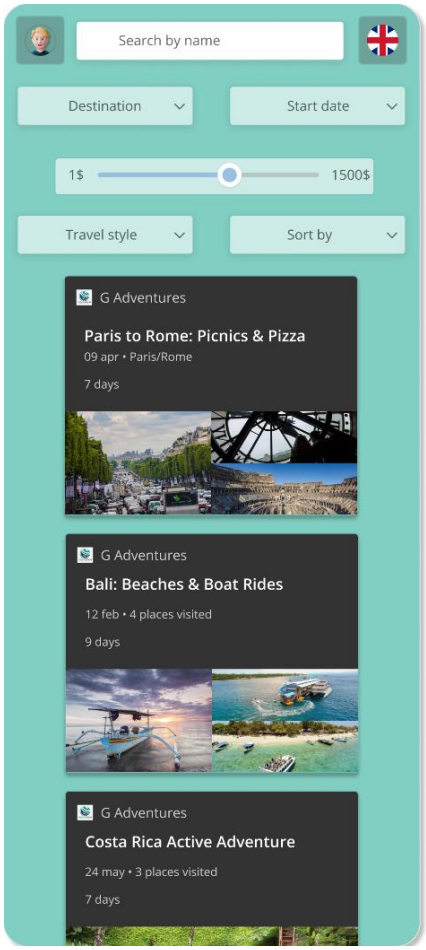


Рисунок 3.10 – Введення даних та вивід рекомендаційних маршрутів

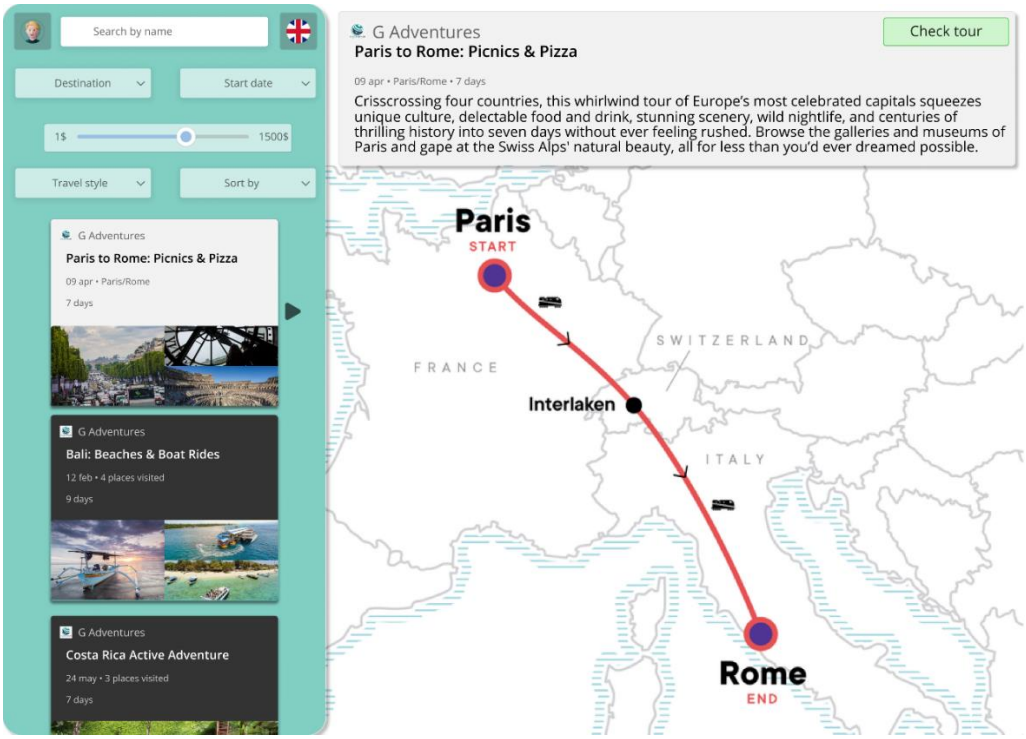


Рисунок 3.11 – Виведення деталей обраного варіанту

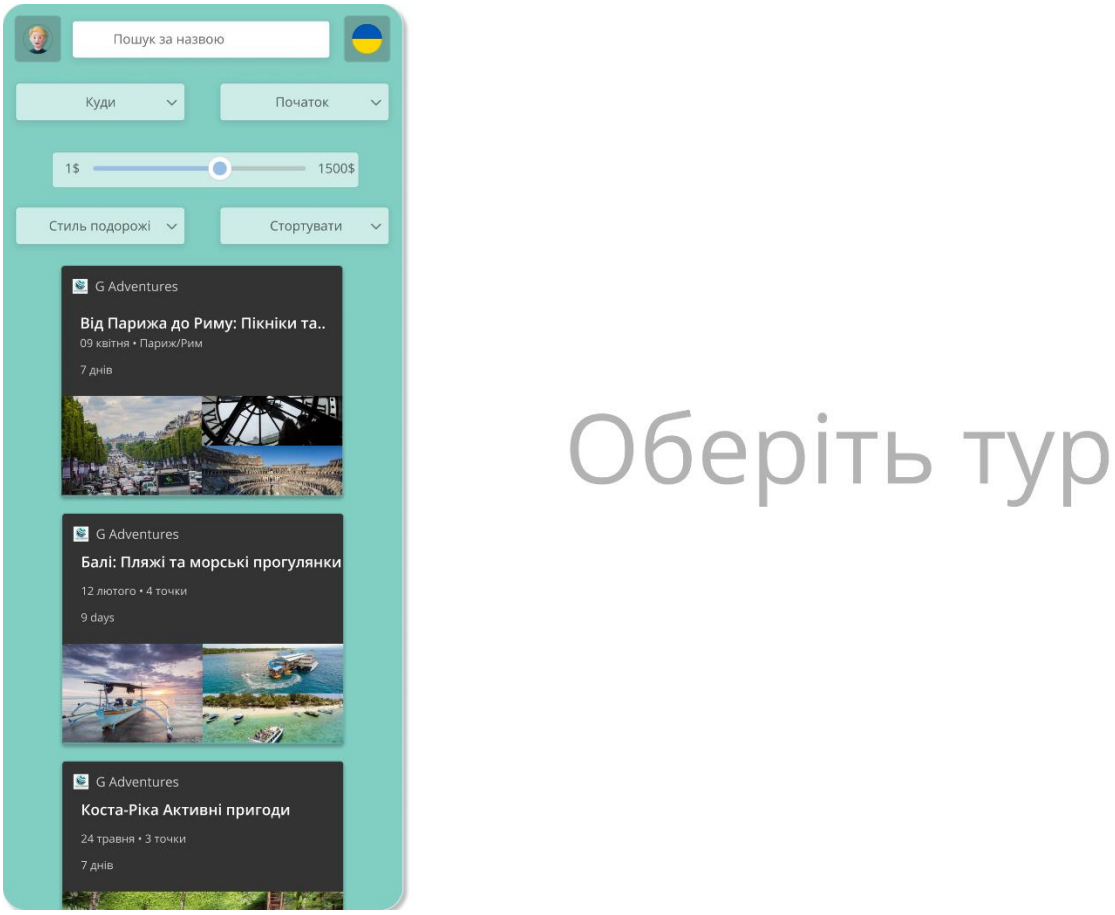


Рисунок 3.12 – Зміна мови в розробленому додатку

В додатку є можливість вибору мови інтерфейсу, що полегшує користування додатком жителю будь-якої країни (рис. 3.12).

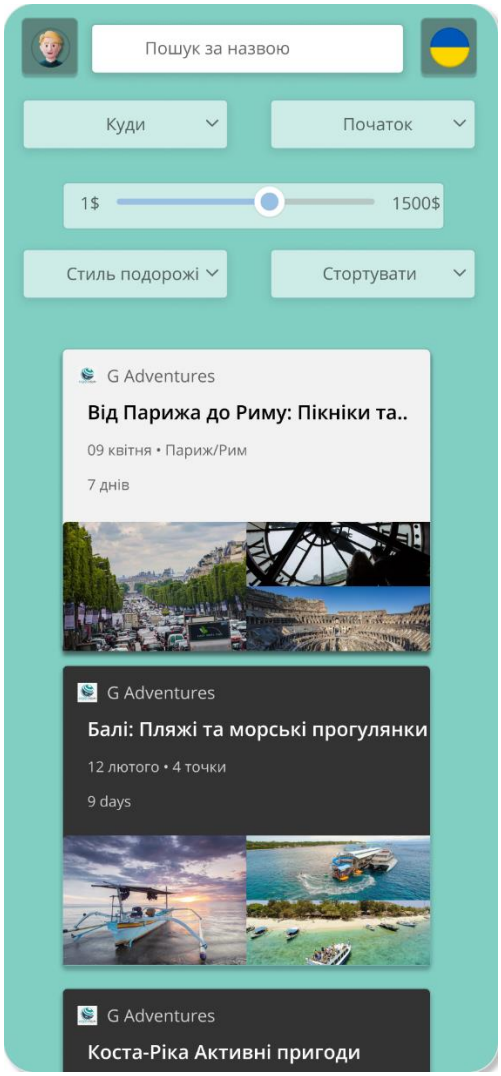


Рисунок 3.13 – Мобільна версія додатку

Отже запропонована інформаційна технологія допомагає оптимізувати роботу туроператора та зменшити час при виборі туристичного маршруту самим туристом.

3.4 Дослідження ефективності запропонованої системи

У домені туристичних рекомендаційних систем найбільш загальновизнаними мірами оцінки ефективності є точність, запам’ятовування та *f*-оцінка. Загалом,

точність і рівень помилок, обчислені на основі тестового набору даних, є основними вимірюваннями, які використовуються для оцінки ефективності моделі. Основною вимогою є модель з найвищим рівнем точності або найменшим рівнем помилок. Однак сама по собі точність або частота помилок не гарантує, що тестова модель працює добре; кілька інших вимірювань також корисні для порівняння продуктивності різних моделей. У задачі багатокласової класифікації модель може отримати пристойний рівень точності, але це може призвести до зниження продуктивності для окремих класів.

Точність є мірою продуктивності класифікатора. Він відображає загальну правильність моделі. Його можна розрахувати як суму правильних класифікацій, поділену на загальну кількість класифікацій, як показано в наступному рівнянні:

$$Accuracy = \frac{|TP| + |TN|}{|FN| + |FP| + |TN| + |TP|} \tag{3.8}$$

Подібним чином продуктивність класифікатора іноді можна виразити в термінах частоти помилок неправильної класифікації. Рівень помилок можна розрахувати за такою формулою:

$$Error rate = \frac{|FN| + |FP|}{|FN| + |FP| + |TN| + |TP|} \tag{3.9}$$

Таблиця 3.1 – Матриця помилок

		Прогноз	
		Клас 1	Клас 0
Фактичне значення	Клас 1	TP	FN
	Клас 0	FP	TN

Матриця помилок або таблиця помилок містить інформацію щодо фактичних і прогнозованих класифікацій, створених класифікатором. Інформація складається

з істинного позитивного (TP), істинного негативного (TN), хибного позитивного (FP) і хибного негативного (FN). У таблиці нижче наведено приклад матриці помилок.

Використання лише точності чи частоти помилок у багатьох випадках може ввести в оману, особливо в реальних проблемах, де набір даних зазвичай незбалансований, як у нашому випадку. Уявіть собі задачу двійкової класифікації, у якій є 900 зразків класу A та 100 зразків класу B. Якби класифікатор передбачив, що все має бути класом A, це дало б високий рівень точності класифікації 90%. Однак класифікатор не може виявити клас B. Для оцінки продуктивності класифікатора використовуються показники точності та релевантності. Точність вказує, скільки вибраних елементів є релевантними. Вимірювання точності та запам’ятовування розраховуються за такими формулами:

$$Precision = \frac{|TP|}{|FP| + |TP|} \tag{3.10}$$

$$Recall = \frac{|TP|}{|FN| + |TP|} \tag{3.11}$$

У домені системи рекомендацій точність важливіша, ніж запам’ятовування, оскільки можна досягти вищої точності, а не запам’ятовування.

F-міра є комбінацією двох вимірювань: точності та запам'ятовування. F-міру можна розглядати як покращення точності, оскільки вона враховує класове розділення. Максимальне значення F-показника – 1, мінімальне – 0. Формула F-показника представлена нижче:

$$Fscore = 2 \times \left(\frac{precision \times recall}{precision + recall} \right) \tag{3.12}$$

Крива робочих характеристик приймача (ROC) – це графік, який представляє продуктивність класифікатора шляхом побудови графіка TP проти FP на кількох

порогових значеннях, як показано на рис. 3.10. Крива ROC використовувалася для порівняння продуктивності кількох моделей машинного навчання та демонструє низку бажаних властивостей у порівнянні з точністю класифікації. Класифікатор, який має криву ROC близько до верхнього лівого кута, вважається кращим за інші. З іншого боку, класифікатор, який має криву ROC нижче діагональної лінії, вважається гіршим. Згідно з рис. 3.10, класифікатор B (ряд 1) вважається кращим (тобто кращим щодо ефективності рекомендацій) порівняно з класифікаторами A (ряд 3) та C (ряд 2).

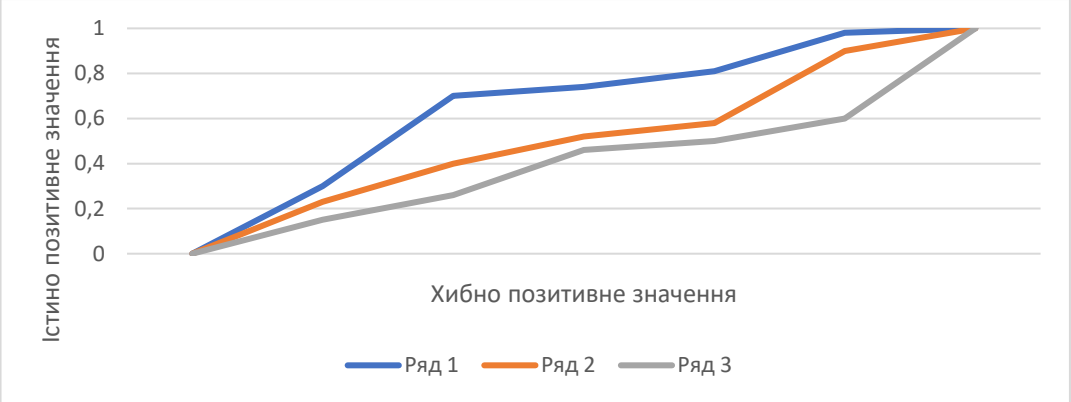


Рисунок 3.10 – Порівняння продуктивності класифікаторів за допомогою кривих ROC

Крива площі під робочими характеристиками приймача, також відома як площа під кривою (AUC), використовується як один із показників для оцінки алгоритму класифікації. AUC можна розрахувати, вимірявши площу під кривою AUC. AUC використовується, щоб визначити, наскільки добре модель класифікації може розрізняти два класи. Чим ближче значення AUC до 1, тим краща модель. Модель, яка має значення AUC, близьке до базового рівня 0,5, вважається марною та не кращою за випадкове припущення.

Щоб вибрати оптимальну модель, оцінити продуктивність моделі та захистити від змін в прогнозній моделі, у цьому дослідженні було використано методи перехресної перевірки. Ці методи застосовуються на етапах регуляризації моделі та оцінки моделі. Перехресна перевірка намагається максимізувати

навчальні дані. Найпростіший підхід для перехресної перевірки починається з двох згорток, у яких набір даних розбивається на дві частини, які називаються навчанням і тестуванням. На наступній ітерації тестовий набір даних міняється місцями з навчальним набором даних.

Цей метод було узагальнено за допомогою *k*-кратної перехресної перевірки, щоб розділити набір даних на *k* розділів приблизно однакового розміру. Для кожної ітерації один згорток/розділ вибирався для перевірки набору даних, а решта вибиралися як навчальний набір даних; цей процес повторювався *k* разів. Найпоширеніша *k*-кратна перехресна перевірка включає 5-кратну та 10-кратну перехресну перевірку. При виборі кількості згинів, чим більше значення *k*, тим менше зміщення та висока дисперсія моделі. Показник точності моделі оцінюється як середнє значення точності *k* моделей. У цьому дослідженні *k* встановлено на 5 для всіх експериментів через обмежену обчислювальну потужність.

Метою використання статистичних тестів у цьому дослідженні є порівняння загальної ефективності різних класифікаторів і оцінка стабільності моделей. Після етапу класифікації застосовуються два статистичні тести. По-перше, тест нормальності Шапіро-Вілка використовувався, щоб перевірити, чи дані були нормально розподілені. Статистичний тест Шапіро Вілка визначається таким чином:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.13}$$

де x_i — найменше число у вибірці, а x — середнє значення вибірок. Константу a_i можна розрахувати наступним чином:

$$(a_1, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}} \tag{3.14}$$

Метод Шапіро-Вілка використовується для вибірок розміром менше 2000. Якщо розмір вибірки перевищує 2000, замість нього застосовується тест Колмогорова-Смирнова. Дані не вважаються нормально розподіленими, якщо значення значущості близьке до нуля (>0,05). Далі, якщо дані розподілені нормально, було проведено парний Т-тест з 95% рівнем довіри, щоб визначити, чи відрізняються середні відмінності між парними зразками більш ніж на 0,5. В іншому випадку застосовувався знаковий ранговий тест.

ВИСНОВКИ

Отже, мета магістерської роботи, яка полягає у покращенні процесу вибору туристичних маршрутів за допомогою інформаційної технології на основі рекомендаційних систем з використанням методів машинного навчання досягнута.

В процесі написання магістерської роботи було проаналізовано рекомендаційні системи та визначені їх особливості. Встановлено обмеження для системи. Визначено необхідність інтелектуалізувати системи за допомогою машинного навчання.

На основі проведеного аналізу обрано методи машинного навчання, які дають найкращий результат по класифікації та прогнозуванню даних. А саме, було запропоновано гібридний підхід на основі ансамблю (дерево рішень, метод опорних векторів і багатошаровий перцептрон) для підвищення ефективності системи на основі моделі з точки зору ефективності класифікації. Результати класифікації, такі як прогнозування, оцінка ймовірності та ранжирування з алгоритмів класифікації, об'єднуються для отримання єдиного та більш надійного кінцевого результату. Ефективність класифікаторів оцінювалася за допомогою восьми наборів даних про вибір місця призначення, які були створені при проведенні дослідження.

Запропонована структура рекомендацій місць призначення складається з п'яти підсистем, що ґрунтуються на процесі інтелектуального аналізу даних: отримання даних, попередня обробка даних, вибір ознак, класифікація та побудова моделі, інтерпретація результатів.

Визначено, що сучасні рекомендаційні системи повинні володіти наступними властивостями: покращення процесу прийняття туристичних рішень; зменшити зусилля користувача; продуктивність, швидкість, точність рекомендацій; інтелектуальний інтерфейс користувача або веб-сайт; інтеграція різнорідної інформації; побудова цілісного плану подорожі; оптимальні рекомендації для групи туристів; високоадаптивна система; забезпечення конфіденційності користувачів.

В досліджені було запропоновано застосовувати декомпозицію класу на етапі попередньої обробки. Мета полягала в тому, щоб ідентифікувати групи напрямків зі спорідненими моделями. Декомпозиція класів дає нам багато переваг, включаючи підвищену продуктивність класифікації, масштабованість до великої бази даних, підвищену зрозумілість, модульність і придатність для паралельних обчислень.

Запропонована інформаційна технологія базується на трирівневій архітектурі веб-моделі, більш відомої як клієнт-серверна архітектура. Архітектура, яка складається з трьох рівнів, складається з рівня презентації, додатків і даних.

В роботі наведено використання статистичних тестів, які дозволяють порівняти загальну ефективності різних класифікаторів і оцінку стабільності моделей.

Напрямками подальшого удосконалення даної роботи є удосконалення запропонованої методики за рахунок збільшення кількості показників на основі яких будується прогноз та виведення в програмному середовищі графічних даних прогнозу основних туристичних напрямків.

ЛІТЕРАТУРА

1. Alex Smola. Introduction to Machine Learning, 234 p., 2008.

2. Alptekin, G.I., Buyukozkan, G., 2011. An integrated case-based reasoning and MCDM system for Web based tourism destination planning. EXPERT SYSTEMS WITH APPLICATIONS 38, 2125–2132.

3. Christopher M Bishop. Pattern recognition. Machine Learning, 128 p., 2006.

4. De Bruyn, A., Liechty, J.C., Huizingh, E.K.R.E., Lilien, G.L., 2008. Offering Online Recommendations with Minimum Customer Input Through Conjoint-Based Decision Aids. Marketing Science 27, 443–460. https://doi.org/10.1287/mksc.1070.0306

5. Ethem Alpaydin. Introduction To Machine Learning, 584 p., 2009.

6. Fouss, F., Saerens, M., 2008. Evaluating Performance of Recommender Systems: An Experimental Comparison, in: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Presented at the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 735–738. https://doi.org/10.1109/WIIAT.2008.252

7. Gretzel, U.(1), Hwang, Y.-H.(2), Fesenmaier, D. r. (3), 2012. Informing destination recommender systems design and evaluation through quantitative research. International Journal of Culture, Tourism, and Hospitality Research 6, 297–315. https://doi.org/10.1108/17506181211265040

8. Kantamneni, A., Brown, L.E., Parker, G., Weaver, W.W., 2015. Survey of multi-agent systems for microgrid control. Engineering Applications of Artificial Intelligence 45, 192–203. https://doi.org/10.1016/j.engappai.2015.07.005

9. Koren, Y., Bell, R., Volinsky, C., 2009. Matrix Factorization Techniques for Recommender Systems. Computer 42, 30–37. https://doi.org/10.1109/MC.2009.263

10. Lucas, J.P., Luz, N., Moreno, M.N., Anacleto, R., Almeida Figueiredo, A., Martins, C., 2013. A hybrid recommendation approach for a tourism system. Expert Systems with Applications 40, 3532–3550. https://doi.org/10.1016/j.eswa.2012.12.061

11. Payr, S., Petta, P., Trappl, R., 2002. Emotions in Humans and Artifacts. MIT Press, Cambridge, Mass.

12. Pitoska, E., 2013. E-Tourism: The Use of Internet and Information and Communication Technologies in Tourism: The Case of Hotel Units in Peripheral Areas. Tourism in South East Europe 2, 335–344.

13. Ricci, F., Rokach, L., Shapira, B., 2011. Introduction to Recommender Systems Handbook, in: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.), Recommender Systems Handbook. Springer US, pp. 1–35.

14. Saleh, E., Błaszczyński, J., Moreno, A., Valls, A., Romero-Aroca, P., de la Riva-Fernández, S., Słowiński, R., 2017. Learning ensemble classifiers for diabetic retinopathy assessment. Artificial Intelligence in Medicine. https://doi.org/10.1016/j.artmed.2017.09.006

15. Stephen Marsland. Machine Learning: An Algorithmic Perspective, 452 p., 2015.

16. Tom M. Mitchell. Machine Learning [http://www.cs.cmu.edu/~tom/mlbook.html]

17. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction, 764 p., 2008.

18. Yaser S. Abu-Mostafa. Learning from data, 215 p., 2017

19. Електронне джерело: https://uk.economy-pedia.com/11036128-stratified-sampling. Дата звернення 28.11.2022.

20. Мелешко Є.В. Дослідження методів побудови рекомендаційних систем заснованих на фільтрації контенту // Збірник тез III Міжнародної науково-практичної конференції «Інформаційна безпека та комп'ютерні технології», м. Кропивницький, 19-20 квітня 2018 р. – Кропивницький: ЦНТУ. – 2018. – С. 234-237.

21. Мелешко Є.В., Хох В.Д. Дослідження моделей рекомендаційних систем на основі прихованих факторів // Збірник тез II Міжнародної

науковопрактичної конференції “Інформаційна безпека та інформаційні технології”, м. Кропивницький, 2-3 квітня 2020 р. – Кропивницький: ЦНТУ. – 2020. – С. 46.

22. Міхав В.В., Мелешко Є.В. Метод оптимізації швидкодії бінарних діаграм рішень при представленні даних рекомендаційної системи // Збірник тез II Міжнародної науково-практичної конференції “Інформаційна безпека та інформаційні технології”, м. Кропивницький, 2-3 квітня 2020 р. – Кропивницький: ЦНТУ. – 2020. – С. 17.

23. Трегубчак І.М. Методика вибору та обробки даних для побудови моделі рекомендаційної системи в сфері туризму // XV Науково-технічна конференція «Сучасні інфокомунікаційні технології» . – Київ: ДУТ, 2022.

24. Трегубчак І.М. Методика вибору туристичних маршрутів з використанням методів машинного навчання // «ТІТ». №5, 2022

ДОДАТОК

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

МАГІСТЕРСЬКА РОБОТА
на тему

«Розробка інформаційної технології для рекомендації туристичних маршрутів на основі методів машинного навчання»

Виконав : Студент групи ПДМ-62 Трегубчак Ілля Михайлович

Керівник: к.е.н., доц. Аверічев І.М.

Київ - 2022

СУЧАСНІ ТУРИСТИЧНІ РЕКОМЕНДАЦІЙНІ СИСТЕМИ

Особливості сучасних туристичних рекомендаційних систем:

1. Зосереджені на рекомендації пунктів призначення та маршрутів
2. Інтелектуальні, інтерактивні, адаптивні та автоматизовані.
3. Ефективні та менш нав'язливі
4. Використовуютьсучасні інформаційні технології.

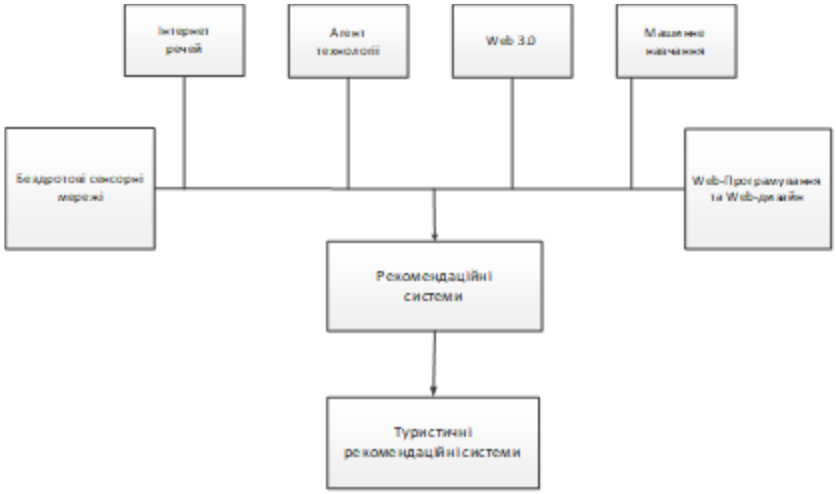


Рис.1. Інформаційні технології в рекомендаційних системах

МЕТА, ОБ'ЄКТА ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета: покращення процесу вибору туристичних маршрутів за допомогою інформаційної технології на основі рекомендаційних систем з використанням методів машинного навчання.

Об'єкт: процес вибору туристичних маршрутів

Предмет: методи машинного навчання та рекомендаційні системи вибору туристичних маршрутів.

ОТРИМАННЯ НЕОБХІДНОГО НАБОРУ ДАНИХ ТА ЇХ ОБРОБКА

Набори даних подорожі:

1. Характеристики подорожі
2. Сума витрат на поїздку.
3. Бажані види діяльності.
4. Мета подорожі
5. Задоволення туристів характеристиками подорожі
6. Демографічна інформація про туристів

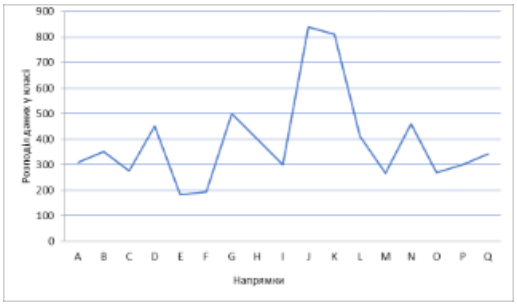


Рис. 1. Розподіл класів для наборів даних

Алгоритм обробки даних:

1. Отримання даних.
2. Попередня обробка даних.
3. Вибір ознак.
4. Класифікація та побудова моделі.
5. Інтерпретація результатів

Мінімально -максимальна нормалізація даних:

$$Normalized(f_i) = \frac{(f - F_{min})}{F_{max} - F_{min}}$$

Нормалізація Z-показника :

$$Normalized(f) = \frac{(f - \bar{f})}{s}$$

АЛГОРИТМИ ОБРОБКИ НАБОРУ ДАНИХ

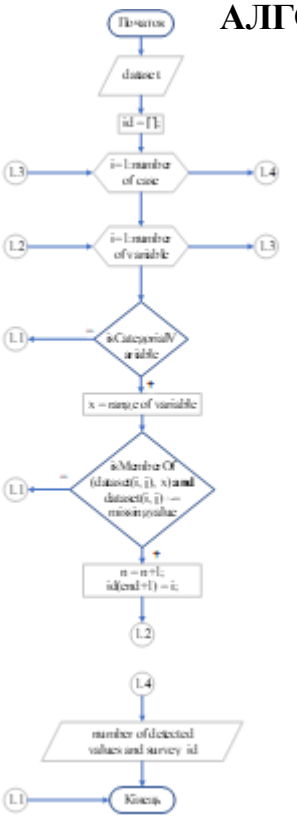


Рис. 1. Алгоритм виявлення викидів

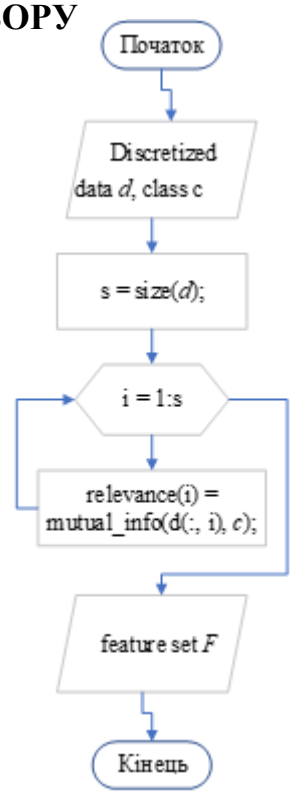


Рис. 2. Алгоритм максимальної релевантності

АЛГОРИТМИ ОБРОБКИ НАБОРУ ДАНИХ

Рис. 1. Алгоритм мінімальної надлишковості та максимальної релевантності

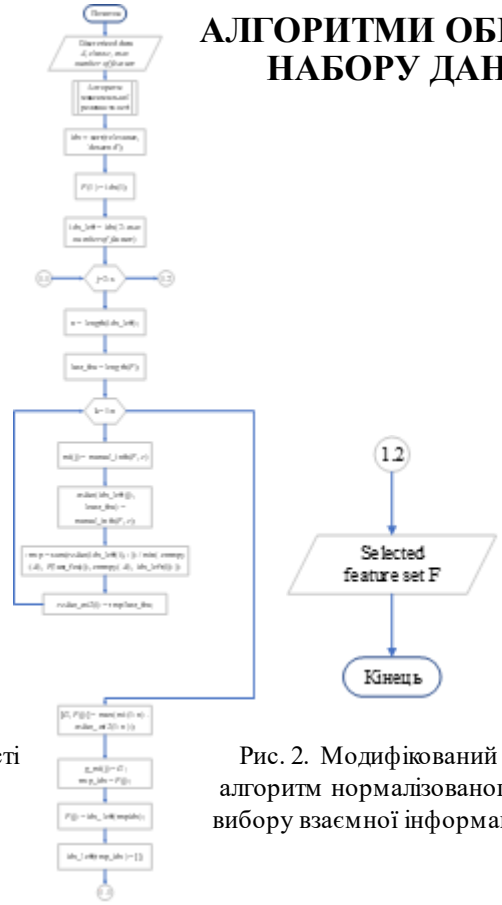


Рис. 2. Модифікований алгоритм нормалізованого вибору взаємної інформації

КЛАСИФІКАЦІЯ ОТРИМАНИХ ДАНИХ

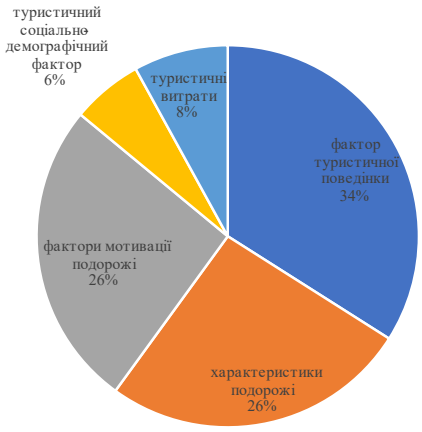


Рис. 1. Розподіл головних факторів, які були використані в моделях вибору місця призначення

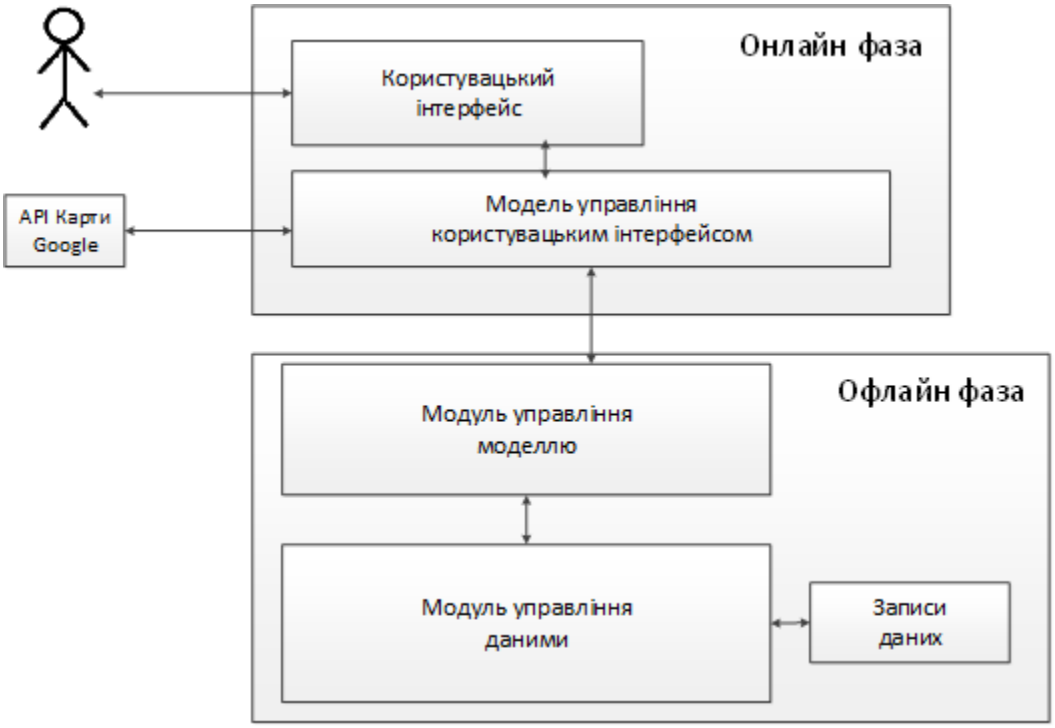
Методи машинного навчання, які використовувалися в роботі для класифікації даних:

1. Дерево рішень .
2. Метод опорних векторів.
3. Багатошаровий персептрон.

Класифікація даних:

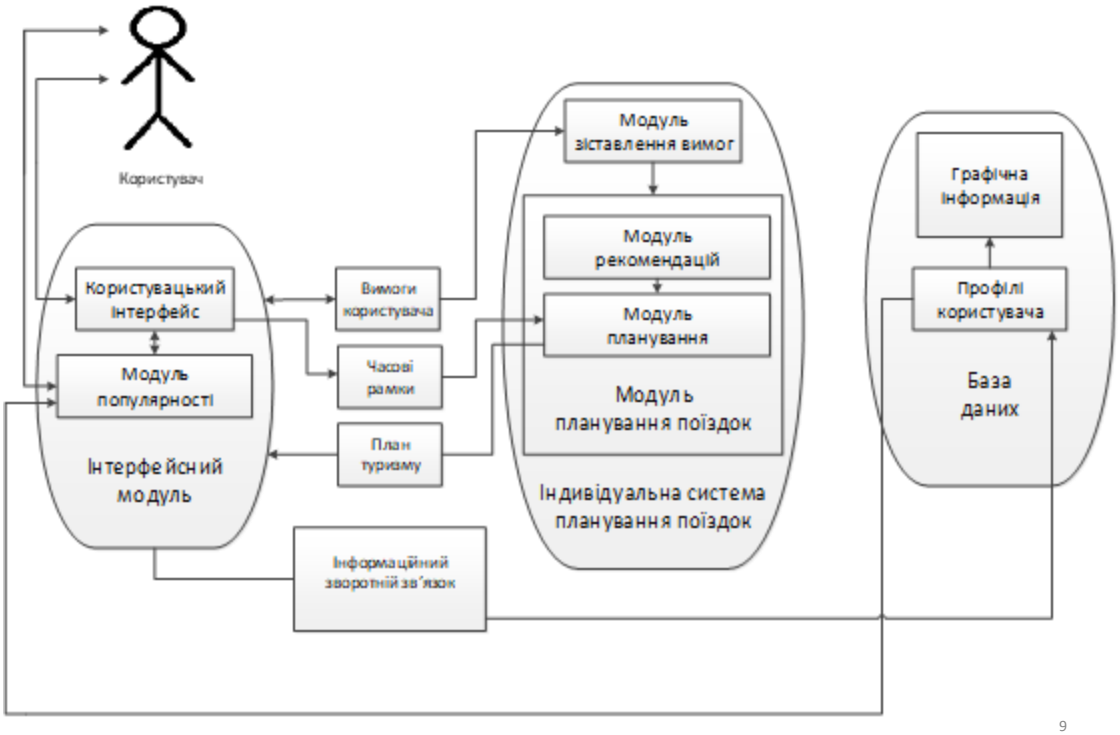
1. Провести масштабування входу та виходу.
2. Використати перехресний пошук оптимальної кількості прихованих нейронів .
3. Навчання мережі з отриманою оптимальною кількістю прихованих нейронів .
4. Тестування з тестовими даними та оцінювання отриманих результатів.

ПРАКТИЧНА СИСТЕМА РЕКОМЕНДАЦІЙ ДЛЯ ТУРИСТІВ



8

ЗАГАЛЬНА СТРУКТУРА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ



9

ЕКРАННІ ФОРМИ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ

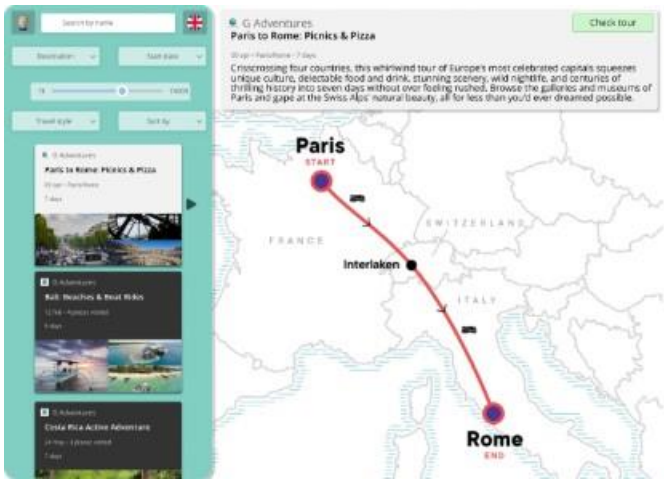


Рис. 1 . Web-додаток на основі запропонованої рекомендаційної системи

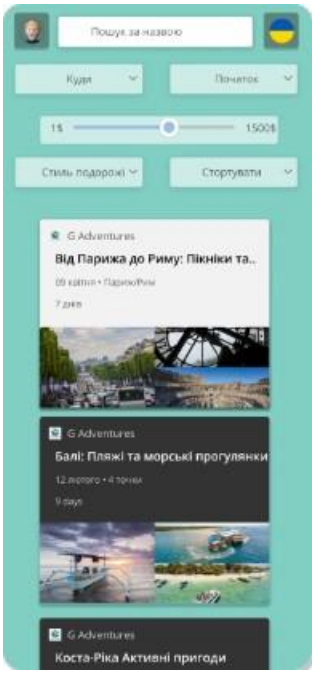


Рис. 1 . Мобільний додаток на основі запропонованої рекомендаційної системи

10

ВИСНОВКИ

9

1. Отже, мета магістерської роботи, яка полягає у покращенні процесу вибору туристичних маршрутів за допомогою інформаційної технології на основі рекомендаційних систем з використанням методів машинного навчання досягнута .
2. На основі проведеного аналізу обрано методи машинного навчання, які дають найкращий результат по класифікації та прогнозуванню даних.
3. Запропонована структура рекомендацій місць призначення складається з п’яти підсистем, що ґрунтуються на процесі інтелектуального аналізу даних: отримання даних, попередня обробка даних, вибір ознак, класифікація та побудова моделі, інтерпретація результатів .
4. В роботі розроблено методику опрацювання вхідних даних для ефективної роботи рекомендаційних систем, яка базується на алгоритмах виявлення викидів, мінімальної надлишковості та максимальної релевантності та модифікованому алгоритмі нормалізованого вибору взаємної інформації
5. Запропонована інформаційна технологія базується на трирівневій архітектурі веб-моделі, більш відомої як клієнт-серверна архітектура . Архітектура, яка складається з трьох рівнів, складається з рівня презентації, додатків і даних.

ПУБЛІКАЦІЯ ТА АПРОБАЦІЯ

Статті:

Трегубчак І.М., Аверічев І.М. Методика вибору туристичних маршрутів з використанням методів машинного навчання // «Зв'язок». №5, 2022, Подана до друку

Тези доповідей

Трегубчак І.М. Методика вибору та обробки даних для побудови моделі рекомендаційної системи в сфері туризму // XV Науково-технічна конференція «Сучасні інфокомунікаційні технології» . – Київ: ДУТ, 2022. – Подана до друку

ДЯКУЮ ЗА УВАГУ!