НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

«До захисту допущено»
В.о. завідувача кафедри

_____ М.В.Грайворонський
(підпис)

“____” _____2019 р.

Дипломна робота

на здобуття ступеня бакалавра

з напряму підготовки 6.040301 «Прикладна математика»

на тему: Метод оцінки росту міста за супутниковими даними

Виконав : студент 4 курсу, групи ФІ-51

Харченко Денис Володимирович
(прізвище, ім’я, по батькові) (підпис)

Керівник к.т.н., доцент Лавренюк Алла Миколаївна
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Консультант _____ _____ _____
(назва розділу) (посада, вчене звання, науковий ступінь, прізвище, ініціали) (підпис)

Рецензент _____ _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ - 2019 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

Рівень вищої освіти – перший (бакалаврський)

Напрям підготовки 6.040301 «Прикладна математика»

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ М.В.Грайворонський
(підпис)

«___»_____2019 р.

ЗАВДАННЯ
на дипломну роботу студенту

Харченку Денису Володимировичу
(прізвище, ім'я, по батькові)

1. Тема роботи: Метод оцінки росту міста за супутниковими даними,
науковий керівник роботи: к.т.н., доцент Лавренюк Алла Миколаївна_____,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «27 » травня 2019 р. № 1414-с

2. Термін подання студентом роботи 10 червня 2019 р.

3. Вихідні дані до роботи: супутникові знімки міста Києва, методи машинного навчання для класифікації супутникових знімків

4. Зміст роботи: провести аналіз досліджень росту міст, описати метод аналізу росту міста, дослідити за допомогою методу ріст міста Києва

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

6. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів дипломної роботи	Примітка
1	Ознайомлення з літературою	05.10.18-19.05.19	
2	Отримання навичок для роботи з ГІС	05.10.18-19.05.19	
3	Збір та аналіз даних	12.02.19-05.05.19	
4	Створення та використання моделі для класифікації супутникових даних	05.05.19-19.05.19	
5	Оформлення дипломної роботи	19.05.19-7.06.19	

Студент _____ _____
(підпис) (ініціали, прізвище)

Керівник роботи _____ _____
(підпис) (ініціали, прізвище)

РЕФЕРАТ

Дипломна робота має обсяг 33 сторінки, 5 рисунків, 4 таблиці, та 9 бібліографічних джерел.

Метою роботи є розробка та дослідження методу машинного навчання для класифікації та оцінки росту міста Києва. Був використаний метод машинного навчання Random Forest.

Об’єкт дослідження:

Ріст міста Києва у 2017 та 2018 роках та його зміна.

Предмет дослідження:

Метод оцінки зміни території штучних об'єктів та зелених насаджень.

В результаті виконання завдання був створений та реалізований метод оцінку росту міста за супутниковими даними для міста Києва.

Ключові слова: ріст міста, супутникові дані, дерева рішень, машинне навчання, класифікація земної поверхні.

ABSTRACT

Thesis paper consists of 33 pages, 5 figures, 4 tables, and 9 bibliographic sources.

The purpose of the work is to develop and study the method of machine learning for the classification and estimation of the Kiev city growth.

Object of study:

The growth of Kyiv city in 2017 and 2018 and its change.

Subject of study:

Method of estimation of changes in the area of artificial objects and green plantations.

As a result of the task, the method urban growth estimation by satellite data was described and applied to the city of Kyiv.

Key words: urban growth, satellite data, decision trees, machine learning, classification of the land cover.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ГІС – геоінформаційна система

SHDI – (Shannon Diversity Index) індекс різноманітності Шеннона

CART - Classification and Regression Tree

NIR – (Near-Infrared) близьке інфрачервоне випромінювання

SWIR – (Short-wavelength infrared) короткохвильове інфрачервоне випромінювання

Confusion Matrix - матриця невідповідностей

ВСТУП

На даний момент ріст міст є гострою проблемою у всьому світі. Тому що збільшення забудівель та зменшення площі зелених насаджень впливає на зміну клімату. Збільшення температури призводить до глобального потепління. Близько 70% вуглекислого газу виробляється у містах, при тому що міста займають менше ніж 2 відсотка площі землі. Збільшення населення у містах призвело до спонтанного та не ефективного росту. З ростом міста збільшується кількість населення та змінюється екологія, що впливає на всіх. Для того щоб контролювати цей процес, потрібно мати актуальні данні про зміни міста. Моніторинг цього явища є непростою задачею під час стрімкого розвитку. Для отримання актуальних даних можна використовувати супутникові дані, що є у відкритому доступі. Для аналізу супутникових даних використовуються методи машинного навчання.

Актуальність роботи:

На сьогоднішній день не існує актуальних даних та оцінок росту міста Києва у відкритому доступі. Також досліджень з використанням супутників високої роздільної здатності що доступні у відкритому доступі для України не має тому задача є актуальною.

Мета і завдання дослідження:

Метою роботи є розробка та дослідження методу машинного навчання для класифікації та оцінки росту міста Києва. Досягнення цієї мети вимагає виконання таких дослідницьких завдань:

1) Аналіз методів застосованих у інших роботах
2) Попередня обробка супутникових знімків
3) Створення вибірки для побудови моделі класифікатора
4) Створення моделі класифікатора земного покрову міста
5) Аналіз результатів класифікації

Об’єкт дослідження:

Ріст міста Києва у 2017 та 2018 роках та його зміна.

Предмет дослідження:

Метод оцінки зміни території штучних об'єктів та зелених насаджень.

Методи дослідження.

1) Класифікація міста методом машинного навчання за супутниковими даними
2) Метод розрахунку показників для оцінки росту міста за результатами класифікації
3) Метод візуального порівняння результатів класифікації та формулювання висновків

Наукова новизна:

Вперше одержано модель класифікатора земного покрову для міста Києва 2017 та 2018 років та проведено аналіз змін земного покрову.

Практична значимість:

Результати класифікації та аналізу міста можуть бути використані у плануванні розвитку міською владою. У роботі можна збільшити кількість років аналіз яких проводився, але це можливо лише у майбутньому коли буде доступно більше даних.

1 ОГЛЯД МЕТОДІВ ОЦІНОК ЗМІНИ ЗЕМНОГО ПОКРОВУ

1.1 Огляд попередніх досліджень

Досліджень на тему росту міст та зміни земного покрову було проведено достатньо у різних країнах та з різними цілями. Але важливими є дослідження що використовували супутникові дані з великим розрізненням, тому що використовуються різні методи. Одне з таких досліджень було проведено у Китаї для провінції Сичуань [7]. У дослідженні використовувались супутникові дані для аналізу земного покрову та його зміни. Використовувався метод машинного навчання Random Forest. Також подібне дослідження проводилось для Риму та Афін у 2013 році, де використовувались супутникові дані для даних регіонів, та карти класифікації, які були аналізовані за допомогою індексів [8]. Для України таких досліджень не проводилось.

1.2 Методологія аналізу земного покрову

В даному розділі розглядається методологія аналізу земного покрову. Для аналізу використовується карта класифікації з заздалегідь визначеними класами. Карт такого типу немає у відкритому доступі, тому вони створюються на основі супутникових даних методами машинного навчання. Машинне навчання - великий підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних навчатися. Для навчання моделі класифікації використовується вибірка, створена вручну за супутниковими знімками.

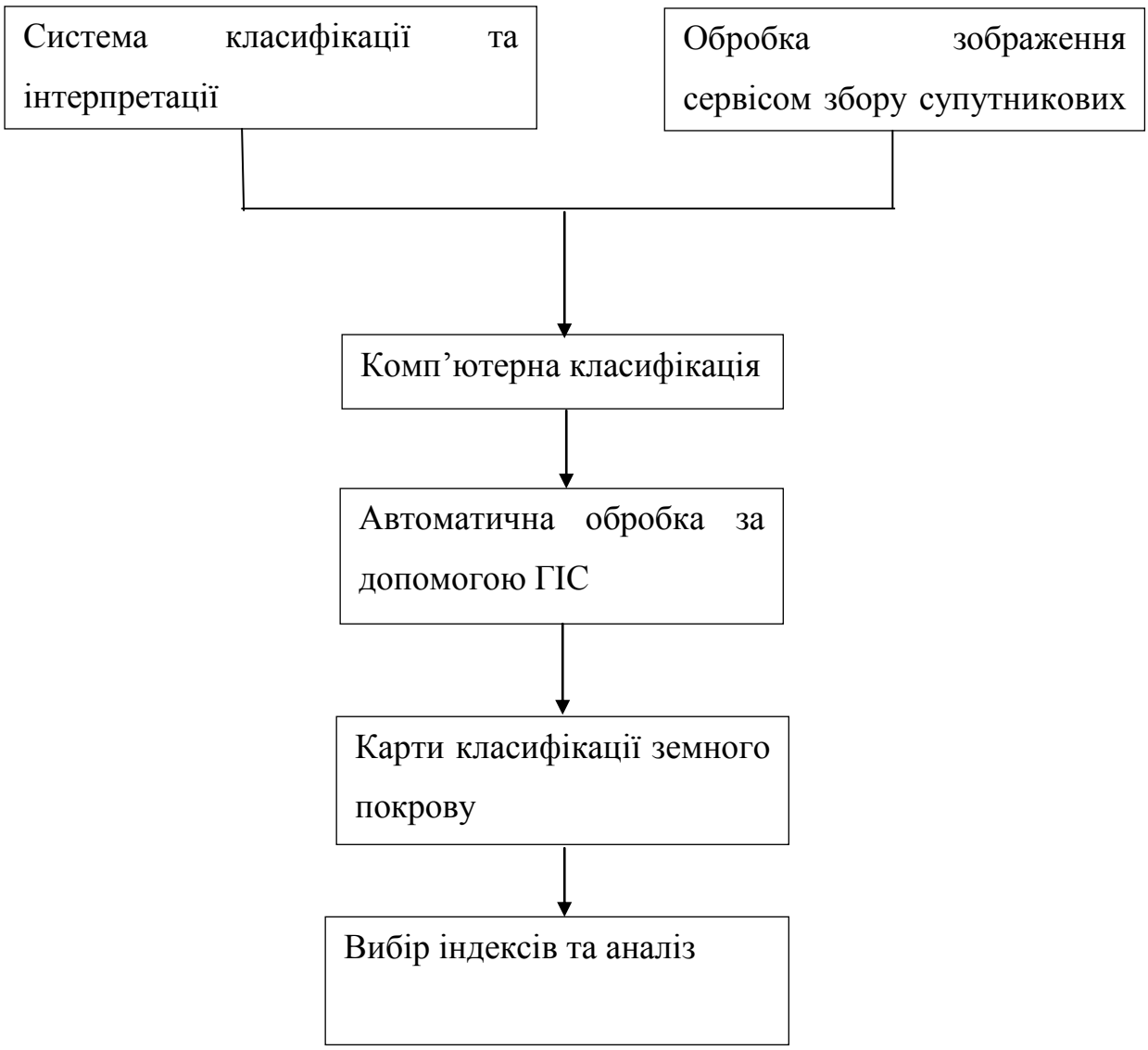

Рисунок 1.1 – Методологія аналізу земного покрову за супутниковими даними

1.3 Класифікація земного покрову методами машинного навчання

В даному розділі розглядаються методи машинного навчання для класифікації земного покрову.

Машинне навчання (Machine Learning) - великий підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних навчатися. Розрізняють два типи навчання. Навчання з вчителем та навчання без вчителя.

Класифікація відноситься до класу задач навчання з учителем. Навчання з учителем - найбільш поширений випадок. Кожен елемент вибірки являє собою пару «об'єкт, відповідь». Потрібно знайти функціональну залежність відповідей від описів об'єктів і побудувати алгоритм, який бере на вході опис об'єкта і видає на виході відповідь. У задачі класифікації множина допустимих відповідей кінцева. Їх називають мітками класів (class label). Клас - це множина всіх об'єктів з даними значенням мітки.

Навчання з учителем полягає в тому, що у вас є вхідні змінні (x) і вихідна змінна (Y), і використовується алгоритм для вивчення відображення

$Y = f(X)$

Мета полягає в тому, щоб наблизити функцію відображення настільки добре, щоб при наявності нових вхідних даних (X) можна було б передбачити вихідні змінні (Y) для цих даних.

Воно називається навчання з учителем, оскільки процес алгоритму навчання з навчального набору може розглядатися як процес навчання вчителем. Відомі правильні відповіді, алгоритм ітеративно робить прогнози на навчальних даних і коригується вчителем. Навчання припиняється, коли алгоритм досягає прийнятного рівня продуктивності.

Одним з методів машинного навчання є дерева рішень. Дерева рішень використовуються для вирішення задача класифікації та регресії.

Дерева рішень - це спосіб представлення правил в ієрархічній, послідовній структурі, де кожному об'єкту відповідає єдиний вузол, що дає рішення.

Алгоритм побудови дерева рішень (CART – Classification and Regression Tree):

Нехай нам задано деяку навчальну множину T, що містить об'єкти (екземпляри), кожен з яких характеризується m атрибутами (ознаками), причому один з них вказує на приналежність об'єкта до певного класу.

Нехай через $\{C_1, C_1, \ldots C_к\}$ позначені класи (значення мітки класу), тоді існують 3 ситуації:

1. Множина T містить один або більше екземплярів, що відносяться до одного класу C_k. Тоді дерево рішень для T - це лист, який визначає клас C_k.
2. Множина T не містить жодного екземпляру, тобто порожня множина. Тоді це знову лист, і клас, асоційований з листом, вибирається з іншої множини відмінної від T, скажімо, з множини, асоційованого з батьком.
3. Множина T містить приклади, що відносяться до різних класів. В цьому випадку слід розбити множину T на деякі підмножини. Для цього вибирається одна з ознак, що має два і більше відмінних один від одного значень $O_1, O_2, \ldots O_n$. T розбивається на підмножини $T_1, T_2, \ldots T_n$, де кожна підмножина T_i містить всі приклади, що мають значення O_i для вибраної ознаки. Це процедура буде рекурсивно тривати до тих пір, поки кінцева множина не буде складатися із екземплярів, що відносяться до одного і того ж класу.

Для побудови дерева на кожному внутрішньому вузлі необхідно знайти таку умову (перевірку), яке б розбивало множину, асоційовану з цим вузлом на підмножини. В якості такої перевірки повинен бути обраний один з атрибутів. Загальне правило для вибору атрибута можна сформулювати наступним чином: обраний атрибут повинен розбити множину так, щоб одержувані в результаті підмножини складалися з об'єктів, що належать до одного класу, або були максимально наближені до цього, тобто кількість об'єктів з інших класів ("домішок") в кожному з цих множин була якомога менше.

Для цього використовується критерій Джинні:

$$Gini(c) = 1 - \sum_j p_j^2 \quad (1.1)$$

де c - поточний вузол, p_j - імовірність класу j в вузлі c .

Random forest [1] - це множина дерев рішень. У задачі регресії їх відповіді усереднюються, в завданні класифікації приймається рішення голосуванням за більшістю.

По суті, Random Forest є композицією (ансамблем) множини дерев рішень, що дозволяє знизити проблему перенавчання і підвищити точність в порівнянні з одним деревом. Прогноз виходить в результаті агрегування відповідей множини дерев. Тренування дерев відбувається незалежно один від одного (на різних підмножинах), що не просто вирішує проблему побудови однакових дерев на одному і тому ж наборі даних, але і робить цей алгоритм досить зручним для застосування в системах розподілених обчислень. Взагалі, ідея беггінга, запропонована Лео Брейманом, добре підходить для розподілу обчислень.

Для беггінга (незалежного навчання алгоритмів класифікації, де результат визначається голосуванням) є сенс використовувати велику кількість дерев рішень з досить великою глибиною. Під час класифікації фінальним результатом буде той клас, за який проголосувало більшість дерев, за умови, що одне дерево має один голос.

Так, наприклад, якщо в завданні бінарної класифікації була сформована модель з 500 деревами, серед яких 100 вказують на нульовий клас, а інші 400 на перший клас, то в результаті модель передбачатиме саме перший клас.

Random Forest (через незалежну побудову глибоких дерев) вимагає дуже багато ресурсів, а обмеження на глибину зашкодить точності (для вирішення складних завдань потрібно побудувати багато глибоких дерев). Можна помітити, що час навчання дерев зростає приблизно лінійно їх кількості.

Природно, збільшення висоти (глибини) дерев не найкращим чином позначається на продуктивності, але підвищує ефективність цього алгоритму

(хоча і разом з цим підвищується схильність до перенавчання). Занадто сильно боятися перенавчання не слід, так як це буде скомпенсоване числом дерев. Але і захоплюватися теж не слід. Скрізь важливі оптимально підібрані параметри (гіперпараметри).

Вибір даного алгоритму не випадковий, на супутникових даних random forest показує найкращі результати с прийнятним часом на виконання. Алгоритм показує кращі результати, ніж SVM [3].

Параметри алгоритму:

1. Найважливішим параметром є кількість дерев на кожний клас. Чим більша кількість дерев, тим кращий результат, але збільшується кількість пам’яті для виконання алгоритму. На практиці кількість дерев має сенс збільшувати лише до певного моменту, потім покращення результату дуже незначне.
2. Кількість змінних для розбиття. Параметр, важливість якого менша, зазвичай, вибирається значення - корінь квадратний з кількості змінних.
3. Мінімальний розмір кінцевого вузла.
4. Частина входу для мішка на одне дерево. Зазвичай береться значення 0,5.

Оцінка точності алгоритму може виконуватися за допомогою матриці невідповідностей (confusion matrix)[9]. Елементом i,j матриці є кількість визначених пікселів для класу i, якщо j - це відомий клас пікселя, тобто на діагоналі буде кількість правильно класифікованих пікселів, тоді як інші елементи показують кількість сплутаних з іншим класом пікселів.

Метод опорних векторів [6] (SVM - support vector machines) - ціль цього методу машинного навчання, знайти гіперплощину у N – мірному просторі, де N – кількість ознак, що розділяє на класи точки. Гіперплощини - це межі рішень, які допомагають класифікувати точки даних. Точки даних, що попадають на будь-яку сторону гіперплощини, можна віднести до різних класів. Крім того, розмір гіперплощини залежить від кількості функцій. Якщо кількість вхідних функцій

дорівнює 2, то гіперплощина - це лише лінія. Якщо число вхідних ознак дорівнює 3, то гіперплощина стає двовимірною площиною. Важко уявити, коли кількість ознак перевищує 3.

Опорні вектори є точками, які ближче до гіперплощини і впливають на положення і орієнтацію гіперплощини. Використовуючи ці опорні вектори, максимізуємо відстань класифікатора. Видалення опорних векторів змінить положення гіперплощини. Це ті пункти, які допомагають побудувати метод опорних векторів.

Метод опорних векторів будує функцію виду:

$$F(x) = sign\big((w, x) + b\big), \tag{1.2}$$

що можна сформулювати як задачу оптимізації:

$$\begin{Bmatrix} \arg min_{w,b}\big||w|\big|^2, \\ y_i\big((w, x_i) + b\big) \geq 1, i = 1, \dots, m. \end{Bmatrix} \tag{1.3}$$

1.4 Супутникові дані

Для класифікації міста використовуються супутникові дані. На даний момент супутникові знімки найкращої роздільної здатності надає програма Європейського космічного агентства Copernicus.

Вона спрямована на досягнення глобального, безперервного, автономного, високоякісного, широкого спектру можливостей спостереження Землі. Забезпечення точної, своєчасної та легкодоступної інформації, серед іншого, для поліпшення управління навколишнім середовищем, розуміння та пом'якшення наслідків зміни клімату, забезпечення цивільної безпеки.

Мета полягає в тому, щоб використовувати величезну кількість глобальних даних із супутників і з наземних, бортових і морських систем вимірювання для

отримання своєчасної та якісної інформації, послуг та знань, а також для забезпечення автономного та незалежного доступу до інформації в сфері навколишнього середовища та безпеки на глобальному рівні, щоб допомогти постачальникам послуг, державним органам та іншим міжнародним організаціям підвищити якість життя громадян Європи. Іншими словами, вона об'єднує всю інформацію, отриману супутниками, повітряними і наземними станціями і датчиками, щоб забезпечити повну картину "здоров'я" Землі.

Важливою частиною програми є місія «Sentinel». Місія «Sentinel» включає радіолокаційну та суперспектральну візуалізацію для моніторингу наземного, океану та атмосферного повітря. Кожна місія «Sentinel» заснована на сузір'ї двох супутників для виконання та перегляду вимог щодо охоплення кожної місії, надання надійних наборів даних для всіх послуг Copernicus.

На даний момент місія включає такі набори:

1. Sentinel-1 забезпечить всепогодний, денний і нічний радіолокаційний станції для наземних і океанічних послуг. Перший супутник Sentinel-1A був успішно запущений 3 квітня 2014 року. Другий супутник Sentinel-1B був запущений 25 квітня 2016 року.
2. Sentinel-2 забезпечить оптичну візуалізацію з високою роздільною здатністю для наземних служб (наприклад, зображення рослинного покриву, ґрунтового та водного покриву, внутрішніх водних шляхів і прибережних районів). Sentinel-2 також надасть інформацію для екстрених служб. Перший супутник Sentinel-2 успішно стартував 23 червня 2015 року.
3. Sentinel-3 буде надавати послуги з моніторингу океану та світу. Перший супутник Sentinel-3A був запущений 16 січня 2016 року;
4. Sentinel-4, розпочатий як корисне навантаження на супутник третього покоління Meteosat, наддасть дані для моніторингу складу атмосфери. Він буде запущений у 2023 році;

5. Sentinel-5 Precursor - це підмножина сенсорного набору Sentinel 5. Він був запущений 13 жовтня 2017 року. Вимірювання проводиться спектроскопом Tropomi.

Для класифікації використовується набір даних с Sentinel-2. Кожен супутник Sentinel-2 несе на собі мультиспектральний прилад з 13-му спектральними каналами у видимому, близькому інфрачервоному (VNIR) і інфрачервоному з короткими хвилями (SWIR) спектральних діапазонах.

Таблиця 1.1 – опис смуг з приладів Sentinel-2

Смуги Sentinel-2	**Центральна довжина хвилі (µм)**	**Роздільна здатність (м)**	**Ширина смуги (нм)**
Смуга 1 — Прибережні аерозолі	0.443	60	20
Смуга 2 — Синій	0.490	10	65
Смуга 3 — Зелений	0.560	10	35
Смуга 4 — Червоний	0.665	10	30
Смуга 5 — Вегетаційний червоний край	0.705	20	15
Смуга 6 — Вегетаційний червоний край	0.740	20	15
Смуга 7 — Вегетаційний червоний край	0.783	20	20
Смуга 8 — NIR	0.842	10	115
Смуга 8A — Вузький NIR	0.865	20	20
Смуга 9 — Водяна пара	0.945	60	20
Смуга 10 — SWIR — Cirrus	1.375	60	20

Продовження таблиці 1.1

Смуги Sentinel-2	**Центральна довжина хвилі (µм)**	**Роздільна здатність (м)**	**Ширина смуги (нм)**
Смуга 11 — SWIR	1.610	20	90
Смуга 12 — SWIR	2.190	20	180

1.5 Методи оцінки змін міста та земного покрову

У дослідженнях інших територій використовуються різні геопросторові показники на основі яких можна робити висновки про зміни та їх характер[4]. Одним з таких показників є показник різноманітності. Найбільш показовим є індекс Шеннона:

$$SHDI = -\sum_{i=1}^{m}(p_i * ln\, p_i) \quad (1.4)$$

Інтерпретація індексу Шеннона – чим більше значення, тим більша кількість інформації, тобто якщо аналізувати місто, то це означає, що воно стало більш складним, а отже розвивається, змінюється та росте.

Також за оцінку росту будемо брати кількість пікселів, яка змінила свій клас з зеленого насадження на штучний об’єкт, що означає забудову зелених зон. Але також треба враховувати кількість пікселів, що змінилась с штучного об’єкта на зелені насадження, що може означати озеленення невикористовуваних територій з бетонним, наприклад, покриттям на парк, що також відповідає зміні на росту міста.

Висновки до розділу 1

У даному розділі було розглянуто методологію аналізу земного покрову та конкретні її етапи. Методом машинного навчання було обрано алгоритм Random Forest, який показує кращі результати за даними попередніх досліджень, що дозволяє отримати потрібні результати класифікації. Супутникові дані використані з комплексу супутників Sentinel-2. А оцінка росту міста проводиться за допомогою індексу різноманітності Шеннона.

2 АНАЛІЗ РОСТУ МІСТА КИЄВА ЗА СУПУТНИКОВИМИ ДАНИМИ

2.1 Опис технічних засобів та програмного забезпечення

Для класифікації земної поверхні міста Києва була використана платформа хмарних обчислень для обробки супутникових даних Google Earth Engine [5].

Earth Engine складається з петабайтних даних, готових до аналізу, спільно розташованих з високопродуктивною, внутрішньо паралельною службою обчислення. Доступ до платформи здійснюється за допомогою інтерфейсу прикладного програмування, доступного в Інтернеті (API), та пов'язаної з ним веб-інтерактивного середовища розробки (IDE), що дозволяє створювати прототипи та візуалізувати результати.

На даній платформі була створена навчальна вибірка за супутниковими знімками 2017-го та 2018-го років комплексу супутників Sentinel-2. Для кожного класу було створено набір полігонів, для класу штучних об’єктів це 65 полігонів, зелених насаджень 38, та води 26. Різна кількість полігонів обумовлена складністю класу та його визначення.

Перед використанням супутникових знімків їх було відфільтровано за кількістю хмар <20% (даний показник було отримано із даних самого знімка), та використано маску для видалення хмар розроблену сервісом, що надає знімки спеціально для цього супутника.

Для подальшого аналізу карт класифікації було використано програмне забезпечення з відкритим початковим кодом Quantum GIS, що представляє собою ГІС систему для роботи с геопросторовими даними.

2.2 Створення моделі класифікатора Random Forest для класифікації міста Києва

За допомогою платформи Google Earth Engine була створена навчальна вибірка для кожного року у місті Києві. На основі цієї вибірки, що складається с 3х класів:

1. Штучні об’єкти – будівлі, дороги, спорудження, створені людиною, які не відносяться до інших класів.
2. Вода – річки, озера.
3. Зелені насадження – дерева, кущі, газони та усі види рослинності.

Вибірка була створена вручну на основі супутникових знімків відфільтрованих за кількістю хмар <20% та медіани за 1 місяць. Використовувались місяці с 4 по 11, тому що в ці місяці немає снігу та можна відрізнити різні об'єкти. Для класу вода були відібрані набори пікселів з великих річок, тому що там найменше всього водоростей на поверхні, тому можна відрізнити зелене насадження від води.

Була створена модель, що приймає на вхід значення каналів супутникового знімка (Таблиця 2.1).

Таблиця 2.1 - Опис ознак, що використовувались для моделі класифікатора

Ім’я	**Розрізнення**	**Довжина хвилі (µм)**	**Опис**
B2	10 метрів	496.6nm (S2A) / 492.1nm (S2B)	Синій
B3	10 метрів	560nm (S2A) / 559nm (S2B)	Зелений
B4	10 метрів	664.5nm (S2A) / 665nm (S2B)	Червоний
B5	20 метрів	703.9nm (S2A) / 703.8nm (S2B)	Вегетаційний червоний край

Продовження таблиці 2.1

Ім’я	Розрізнення	Довжина хвилі (µм)	Опис
B6	20 метрів	740.2nm (S2A) / 739.1nm (S2B)	Вегетаційний червоний край
B7	20 метрів	782.5nm (S2A) / 779.7nm (S2B)	Вегетаційний червоний край
B8	10 метрів	835.1nm (S2A) / 833nm (S2B)	NIR
B8A	20 метрів	864.8nm (S2A) / 864nm (S2B)	Вегетаційний червоний край
B11	20 метрів	1613.7nm (S2A) / 1610.4nm (S2B)	SWIR 1
B12	20 метрів	2202.4nm (S2A) / 2185.7nm (S2B)	SWIR 2

Отже, модель приймає множину ознак та видає відповідь у вигляді мітки класу:

Таблиця 2.2 – Цифрове кодування міток класів

Штучні об’єкти	Вода	Зелені насадження
0	1	2

Модель для класифікації 17-го року була створена з параметрами:

1. 10 дерев на кожний клас.
2. Кількість змінних на розбиття залишена за замовчуванням тобто $\sqrt{M}$, де M – кількість змінних.
3. Мінімальний розмір кінцевого вузла: 1.
4. Частина входу на кожне дерево 0.5.

Для такої моделі :

$$confusion\ matrix\ = \begin{bmatrix} 1889 & 1 & 7 \\ 4 & 2905 & 0 \\ 7 & 0 & 26167 \end{bmatrix}$$

Точність = 0.9993867010974823

Для моделі с кількістю дерев 2:

$$confusion\ matrix\ = \begin{bmatrix} 1888 & 1 & 8 \\ 24 & 2884 & 0 \\ 115 & 3 & 26056 \end{bmatrix}$$

Клас зелених насаджень та штучних об’єктів плутаються достатньо сильно у порівнянні з іншими.

Точність моделі = 0.995093608779858

Збільшення кількості дерев до 10 поліпшує якість класифікації, але подальше збільшення майже не впливає на результат.

Для 15 дерев:

$$confusion\ matrix\ = \begin{bmatrix} 1889 & 1 & 7 \\ 2 & 2907 & 0 \\ 5 & 0 & 26169 \end{bmatrix}$$

Якість класифікації зросла, але кількість дерев збільшилась у півтора рази, що збільшило час та ресурси на обчислення.

Порівняння с алгоритмом CART:

$$confusion\ matrix\ = \begin{bmatrix} 1849 & 15 & 33 \\ 8 & 2902 & 0 \\ 82 & 8 & 26082 \end{bmatrix}$$

Модель класифікатора за цим алгоритмом сплутала значно більш класів, ніж ансамбль дерев класифікатора Random Forest.

Порівняння с алгоритмом SVM:

$$confusion\ matrix\ = \begin{bmatrix} 1847 & 1 & 49 \\ 36 & 2872 & 1 \\ 44 & 0 & 26130 \end{bmatrix}$$

Алгоритм погано класифікував воду, та в цілому показав гірші результати.

Модель для класифікації 18-го року була створена з параметрами:

1. 11 дерев на кожний клас
2. Кількість змінних на розбиття залишена за замовчуванням тобто $\sqrt{M}$ де M – кількість змінних.
3. Мінімальний розмір кінцевого вузла: 1.
4. Частина входу на кожне дерево 0.5.

Для такої моделі :

$$confusion\ matrix\ = \begin{bmatrix} 2139 & 0 & 39 \\ 8 & 5539 & 0 \\ 9 & 0 & 73170 \end{bmatrix}$$

Точність 0.9993078216157422

Для моделі с кількістю дерев 2:

$$confusion\ matrix\ = \begin{bmatrix} 2142 & 0 & 36 \\ 21 & 5526 & 0 \\ 262 & 1 & 72916 \end{bmatrix}$$

Клас зелених насаджень та штучних об’єктів плутаються достатньо сильно у порівнянні з іншими.

Аналогічно з моделлю для 17-го року, для 15 дерев:

$$confusion\ matrix\ = \begin{bmatrix} 2140 & 0 & 38 \\ 6 & 5541 & 0 \\ 9 & 0 & 73170 \end{bmatrix}$$

Точність 0.999344902600613

Точність незначно зросла, але ресурси на обробку виросли значно, тому збільшення кількості дерев далі не дає результату.

Порівняння с алгоритмом CART:

$$confusion\ matrix\ = \begin{bmatrix} 2063 & 2 & 133 \\ 5 & 5541 & 1 \\ 37 & 0 & 73142 \end{bmatrix}$$

Модель з 1 деревом плутає клас штучних об’єктів та зелених насаджень.

Порівняння с алгоритмом SVM:

$$confusion\ matrix\ = \begin{bmatrix} 2050 & 3 & 145 \\ 5 & 5541 & 1 \\ 127 & 4 & 73048 \end{bmatrix}$$

Результат класифікації гірше, кількість неправильно визначених класів більше.

Результат класифікації представлений у вигляді растрової карти (зелений – зелені насадження, червоний – штучні об’єкти, синій – вода)і:

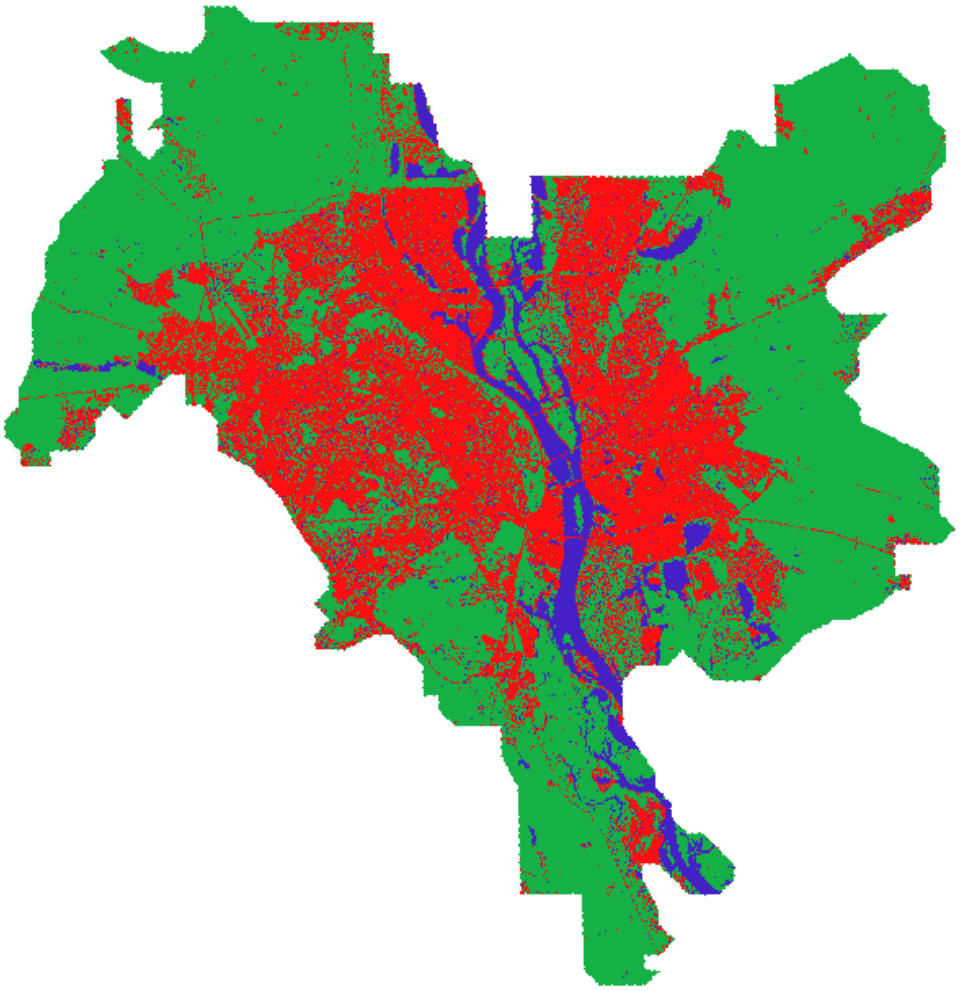

Рисунок 2.1 – Карта класифікації для міста Києва за 2017 рік

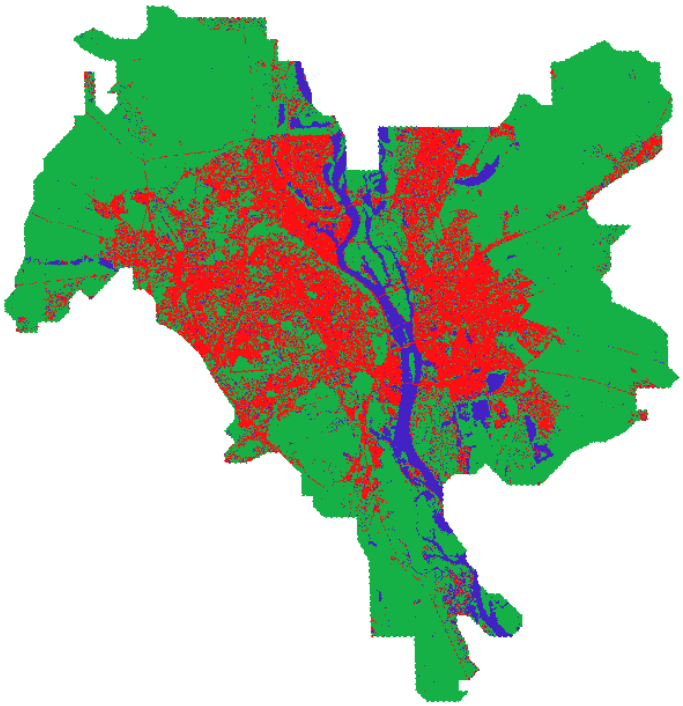

Рисунок 2.2 – Карта Карта класифікації для міста Києва за 2018 рік

2.3 Аналіз отриманих даних

Далі за допомогою програми Qgis можна проаналізувати та оцінити ріст міста.

Спочатку візуальне порівняння, зміни пікселів зелених насаджень на штучні об’єкти:

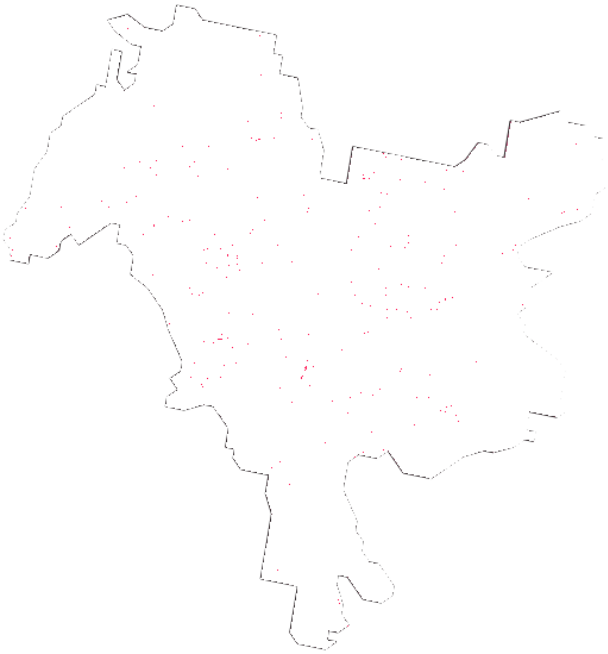

Рисунок 2.3 – Карта з точками, де зелені насадження змінились на штучні об’єкти

Зміна штучних об’єктів на зелені насадження:

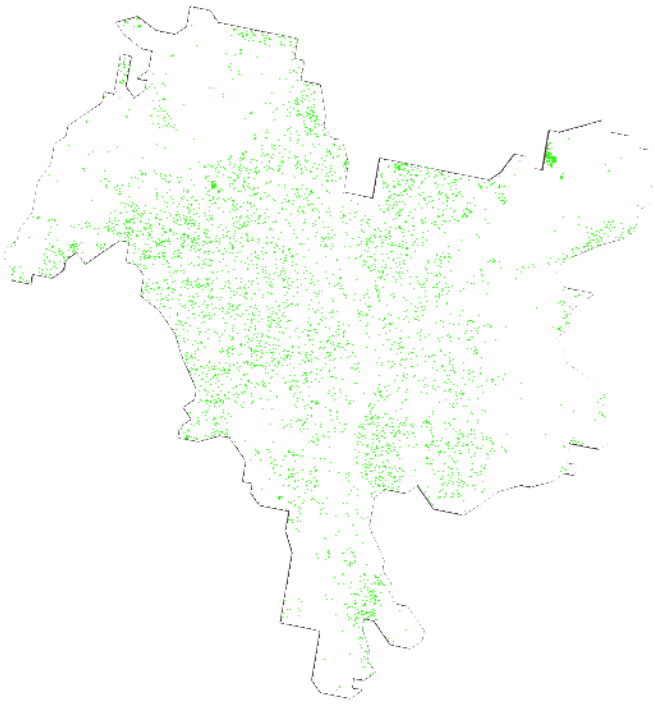

Рисунок 2.4 – Карта з точками, які змінились з штучних об’єктів на зелені насадження

У відсотковому відношенні можна виразити, що загальна площа зелених насаджень зросла на 3,34%, водночас площа штучних об’єктів зменшилась на 2,96%.

Якщо порахувати індекс різноманітності Шеннона:

Таблиця 2.3 - індекс різноманітності Шеннона для 2017 та 2018 років

SHDI для 2017 року	SHDI для 2018 року
0,76393882156869	0,7942614123922926

Індекс різноманітності Шеннона за рік виріс на 0,0303225908236026, це показує, що місто стало більш складним, змінилась забудова, збільшилась кількість зелених насаджень.

Висновки до розділу 2

У розділі розроблено та досліджено метод оцінки росту міста, одним з етапів якого є класифікація земної поверхні, модель класифікатора земної поверхні. Для створеної моделі класифікатора було проведено попередню підготовку та обробку супутникових знімків.

Було досліджено модель класифікатора (Random Forest), яка була порівняна з іншими (CART, SVM), проведено аналіз параметрів моделі та визначено найкращі для виконання поставленої задачі. На основі результатів роботи моделі класифікатора були розраховані матриці невідповідності та значення точності. Розраховано індекс різноманітності Шеннона на основі результату роботи моделі класифікатора. Проведено порівняльний аналіз міста Києва за 2017 та 2018 роки, отримано результати зміни міста, а саме зміни зелених насаджень на штучні об’єкти, та зміни штучних об’єктів на зелені насадження . Загальна площа території зелених насаджень зросла на 3,34% , водночас площа штучних об’єктів зменшилась на 2,96%.

ВИСНОВКИ

У роботі був запропонований та досліджений метод оцінки росту міста за супутниковими даними. Метод був реалізований та застосований для міста Києва. Для методу оцінки була розроблена та досліджена модель класифікатора Random Forest, проведено аналіз та обґрунтування використання даного методу. Оцінка росту міста виконується за допомогою індексу різноманітності Шеннона та відносного порівняння кількості класів. Супутникові дані використовувались з супутника Sentinel-2. Для розробки моделі була проведена попередня обробка супутникових даних, супутникові знімки були відфільтровані за відсотком хмар <20%, та використана маска для усунення тіней від хмар та самих хмар, яка розроблена сервісом, що надає знімки. Були вибрані найбільш придатні для даної задачі ознаки, що дозволили класифікувати земну поверхню міста Києва. Отримана висока точність моделі підтверджується порівнянням з реальними даними.

На основі результатів класифікації було розраховано індекс Шеннона та проведено його аналіз. Результат аналізу показав, що у 2018 році порівняно з 2017 збільшилась кількість штучних об’єктів і, водночас, зросла кількість зелених насаджень, але загальна площа для зелених насаджень зросла на 3,34% , а штучних об’єктів зменшилась на 2,96%.

Карт класифікації для міста Києва не існує у загальному доступі, у порівнянні з містами інших країн. Розроблений метод оцінки росту міста, що використовує відкриті дані та відкрите програмне забезпечення, дозволяє отримати дані, що можуть в подальшому використовуватись органами місцевої влади для планування розвитку міста.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. *LEO BREIMAN* Random Forests // Machine Learning, 45, 5–32, 2001c Kluwer Academic Publishers.

2. *Breiman, L.* Bagging predictors.Machine Learning 26(2), 123–140 (1996a)

3. *Dongping Ming* Land cover classification using random forest with genetic algorithm-based parameter optimization // Journal of Applied Remote Sensing 10(3):035021

4. *Fei Zhang* Assessment of Land-Cover/Land-Use Change and Landscape Patterns in the Two National Nature Reserves of Ebinur LakeWatershed, Xinjiang, China / Hsiang-te Kung and Verner Carl Johnson // Sustainability 2017, 9, 724; doi:10.3390/su9050724

5. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. // Remote Sensing of Environment.

6. *Nello Cristianini, John Shawe-Taylor.* / An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. // Cambridge University Press, 2000. — ISBN 978-1-139-64363-4

7. *Yongjiu Feng , Zongbo Cai , Xiaohua Tong , Jiafeng Wang , Chen Gao , Shurui Chen, Zhenkun Lei* / Urban Growth Modeling and Future Scenario Projection Using Cellular Automata (CA) Models and the R Package Optimx // ISPRS Int. J. Geo-Inf. 2018, 7, 387; doi:10.3390/ijgi7100387

8. *Luca Salvati and Margherita Carlucci* / Urban Growth and Land-Use Structure in Two Mediterranean Regions: An Exploratory Spatial Data Analysis // SAGE Open 2014 4: DOI: 10.1177/2158244014561199

9. *Stephen V. Stehman* / Selecting and Interpreting Measures of Thematic Classification Accuracy // Remote Sensing of Environment 62(1):77-89 · October 1997