МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет радіоелектроніки
Факультет Центр післядипломної освіти
Кафедра Програмної Інженерії

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

Дослідження методів прогнозування дорожнього трафіку

Виконала:
студентка 2 курсу, групи ІПЗздм-21-1
Клочко О.Ю.
(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного забезпечення

Тип програми Освітньо-наукова

Керівник проф. Смеляков С.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. Кафедри _____ З.В. Дудар

2023 р.

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти

Кафедра Програмної Інженерії

Рівень вищої другий (магістерський)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва)

Тип програми освітньо-наукова програма

Освітня програма Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)

« ____» _____20___р

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента Клочко Ольги Юріївни
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження методів прогнозування дорожнього трафіку»

затверджена наказом по університету від 3 квітня 2023 р. №83Стз

2. Термін подання студентом роботи до екзаменаційної комісії 16 травня 2023 р.

3. Вихідні дані до роботи методи регресійного прогнозування, датасет, випадковий ліс, k-найближчих сусідів, градієнтний бустинг, методи оцінки роботи методів, Python, NumPy, Sklearn

4. Перелік питань, що потрібно опрацювати в роботі аналіз предметної області, постановка завдання, аналіз існуючих методів для прогнозування трафіку, формування вимог, підготовка та проведення експерименту, аналіз отриманих результатів

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапівроботи	Примітка
1	Аналіз предметної області	25.02.2023	виконано
2	Постановка завдання	05.03.2023	виконано
3	Аналіз існуючих методів	13.04.2023	виконано
4	Планування дослідження	19.04.2023	виконано
5	Проведення дослідження	10.04.2023	виконано
6	Підготовка пояснювальної записки	30.04.2023	виконано
7	Підготовка презентації та доповіді	02.05.2023	виконано
8	Перевірка на плагіат	07.05.2023	виконано
9	Нормоконтроль	08.05.2023	виконано
10	Рецензування	12.05.2023	виконано
11	Занесення роботи в електронний архів	13.05.2023	виконано
12	Попередній захист	13.05.2023	виконано
13	Допуск до захисту у зав. кафедри	13.05.2023	виконано

Дата видачі завдання ____23 січня 2023 р._____

Студент _____

(підпис)

Керівник роботи _____ проф. Смеляков С.В.

(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Кваліфікаційна робота магістра містить: 70 сторінок, 12 рисунків, 10 таблиць, 32 джерела.

КВАНТИЛЬНА РЕГРЕСІЯ, МАШИННЕ НАВЧАННЯ, ТРАНСПОРТНИЙ ПОТІК, ТРАФІК, ПРОГНОЗУВАННЯ, ШВИДКІСТЬ.

Об’єктом дослідження є методи прогнозування дорожнього трафіку.

Метою роботи є дослідження та аналіз різних методів регресії для прогнозування швидкості транспортного потоку та виявлення найбільш доцільного.

Методи розробки базуються на методах регресійного аналізі, мові програмування Python та її бібліотек.

У результаті був проведений аналіз предметної галузі, поставлені завдання дослідження, аналіз існуючих методів та алгоритмів, планування та проведення дослідження та виявлено найбільш влучний метод для прогнозування швидкості автомобілів.

QUANTILE REGRESSION, MACHINE LEARNING, TRAFFIC FLOW, TRAFFIC, FORECASTING, SPEED.

The object of study is traffic forecasting methods.

The purpose of the study is to investigate and analyze various regression methods for predicting traffic speed and identify the most appropriate one.

The development methods are based on regression analysis methods, the Python programming language and its libraries.

As a result, we analyzed the subject area, set research objectives, analyzed existing methods and algorithms, planned and conducted the study, and identified the most appropriate method for predicting car speed.

Я, Клочко Ольга Юріївна, студентка гр. ІПЗздм-21-1, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження методів прогнозування дорожнього трафіку», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлена з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ВСТУП

Майже всі міста світу стикаються з серйозними проблемами заторів. Надмірний транспортний потік щодня призводить до паралічу міської транспортної системи, що створює великі незручності та негативно впливає на пересування людей. Різні країни активно вживають відповідних заходів, тобто перенаправляють транспортні потоки, обмежують кількість або розширюють масштаби дорожньої мережі, але ці заходи можуть мати незначний ефект [1].

Для управління транспортними потоками використовуються інтелектуальні транспортні системи, які дозволяють в режимі реального часу збирати та обробляти інформацію про дорожню мережу, включаючи швидкість руху, кількість транспортних засобів за певний період, щільність руху, завантаженість дорожньої мережі та розклад руху громадського транспорту.

Необхідність регулювання транспортних потоків у містах України зумовлена кількома причинами [2]: зростанням урбанізації, збільшенням перевантаженості дорожньої мережі, низькою якістю послуг громадського транспорту, незручними маршрутами, тривалим часом у дорозі тощо. Ці проблеми особливо гостро стоять у найбільших містах і спонукають громадян все частіше обирати автомобіль для щоденних поїздок, що, в свою чергу, збільшує затримки, час у дорозі та призводить до забруднення навколишнього середовища [3].

Актуальність цієї роботи полягає в пошуку інструментів для управління та моніторингу цих процесів у містах. Метою роботи є визначення ефективності методів для короткострокового прогнозування інтервалів на основі історичних даних на прикладі середньої швидкості руху автомобілів на годину на конкретній ділянці дороги та оцінка точність прогнозу.

Швидкість руху безпосередньо впливає на реалізацію стратегій управління дорожнім рухом, таких як система керування дорожнім рухом і система орієнтування руху. На точність прогнозування швидкості руху значною мірою впливають доступні дані отримані від радарів і камер

дорожнього руху, встановлених на деяких важливих дорогах. Однак із збільшенням обсягу доступних даних, зібраних із мобільних служб (смартфонів і бортових пристроїв навігації), транспортних засобів-зондів, мікрохвильових датчиків дистанційного руху і різноманітних датчиків Інтернету речей, проблема більше не пов’язана до кількості даних, а радше до вилучення та моделювання корисної інформації з цих даних.

Розвиток заторів на дорогах можна пом’якшити спільно, а умови руху можуть стати більш стабільними. Проте завжди складно реально оцінити короткострокові майбутні умови швидкості руху через складність дорожньої мережі, нестабільність і стохастичність транспортного потоку та швидкість плаваючих транспортних засобів.

Враховуючи мінливість швидкості подорожей за періодичних і одноразових умов руху, мета поточного дослідження полягає в тому, щоб зробити кращі прогнози швидкості. Дані про швидкість були взяти з відкритого сервісу Uber Movement у місті Київ. Було поставлено задачу дослідити короткострокове прогнозування інтервалів швидкості за допомогою квантильної регресії для різних алгоритмів: k-найближчих сусідів, випадкового лісу та градієнтного бустингу. У роботі наведено аналіз отриманих результатів, де загалом моделі дали схожі результати, кардинально відрізнявся тільки час виконання прогнозування.

Проведений аналіз в подальшому може бути використаний для покращення систем дорожнього руху, розробки ефективних стратегій управління дорожнім рухом, таких як оптимальне використання світлофорів та планування дорожніх робіт в часі, коли транспортний потік менший.

Результати роботи представлені на VII Міжнародній конференції «Комп’ютерна лінгвістика та інтелектуальні системи» (CoLInS 2023), що індексується в наукометричих базах CEUR та SCOPUS.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Аналіз предметної галузі

Для управління транспортними потоками використовуються інтелектуальні транспортні системи, які дозволяють в режимі реального часу збирати дані та обробляти інформацію про дорожню мережу, включаючи швидкість руху, кількість транспортних засобів за певний період, щільність потоків, зайнятість дорожньої мережі, розклад громадського транспорту [1].

Міська мобільність є інструментом задоволення потреб громадянина між його функціональними зонами для реалізації зв'язків, що утворюються в результаті життєдіяльності, за допомогою систем транспортного обслуговування або пішки. Основні зв'язки можна поділити на трудові, оздоровчі та культурно-побутові. Система транспортного обслуговування населення міста поділяється на підсистеми громадського та індивідуального транспорту.

Сучасні міста характеризуються інтенсивністю економічних зв'язків, а потреба в транспортних переміщеннях населення настільки велика, що потенційно може бути реалізована лише за умови всебічного розвитку різних видів транспорту і транспортних комунікацій.

Можна виділити такі групи проблем міських транспортних систем [2]:

– задоволення транспортних потреб;
– підвищення економічної ефективності;
– підвищення безпеки руху;
– зменшення шкідливого впливу на навколишнє середовище.

Особливої уваги заслуговують дві проблеми, які найбільше характерні для найбільших українських міст:

– висока залежність населення від автомобіля (коефіцієнт використання автомобіля – 0,88, у містах Європи з розвиненою транспортною інфраструктурою – 0,3);
– завантаженість міст, особливо центрів, приватними автомобілями (Київ займає 3 місце в рейтингу міст з найбільшими затримками [2]).

Аналізуючи досвід багатьох країн світу, можна сформувати два шляхи вирішення проблем транспортних систем міст:

– екстенсивний (збільшення % вулиць і доріг до загальної площі міста, розширення існуючих вул. та дорожньої мережі, будівництво перехресть у різних рівнях тощо);
– інтенсивні (зміна пріоритетів у піраміді міської мобільності (див. рис. 1.1), створення нових маршрутів громадського транспорту та оптимізація існуючих, моделювання транспортних потоків тощо).

Залежно від способу пересування – шлях (піший або транспортний), яким людина добирається від початкового до кінцевого пункту пересування, мобільність може бути загальним або транспортним.

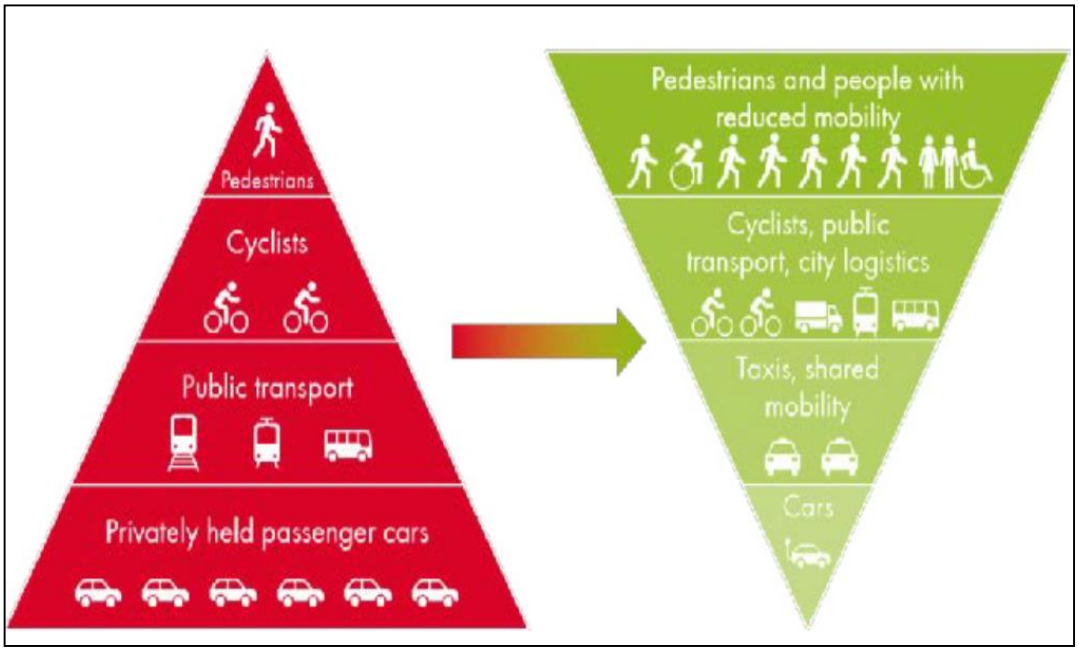

Рисунок 1.1 – Зміна пріоритетів міської мобільності в містах з інтенсивним розвитком транспортних систем [2]

Міста з розвиненою транспортною інфраструктурою при плануванні транспортної інфраструктури надають перевагу пішоходам та маломобільним людям. Другим пріоритетом є громадський транспорт, який може пересувати набагато більше населення, ніж приватні автомобілі, який не вимагає паркування та має набагато менший вплив на навколишнє середовище.

Велосипедний транспорт, який потребує спеціальної інфраструктури та паркувальних місць, також є другим пріоритетом [1].

У багатьох європейських містах інтелектуальні транспортні системи (ІТС) використовуються для моніторингу та управління міською мобільністю. ІТС – це поєднання інновацій у сфері комп'ютерів, інформаційних технологій і телекомунікацій разом із знаннями в автомобільній та транспортній сферах [1]. Саме на основі основних розробок у цих сферах з'являються ключові ІТС-технології. Німецька асоціація міжнародного співробітництва визначає ІТС як «застосування комп'ютерних, інформаційних і комунікаційних технологій для керування транспортними засобами та мережами в режимі реального часу, включаючи рух товарів і людей [1]».

В Україні практично відсутня нормативна база у сфері моніторингу та управління міською мобільністю. У 2018 році в державних будівельних нормах України відбулися важливі зміни, після яких у нормативній базі з'явилися такі поняття, як «транспортне моделювання», «транспортні розрахунки», «кількість переміщень населення» та «імітаційні транспортні моделі». Саме ці зміни дозволили змоделювати ситуацію на всіх елементах вулично-дорожньої мережі міста за допомогою сучасного програмного забезпечення [3]. Водночас варто зазначити, що українське законодавство поки що не має вимог до виконання самого транспортного моделювання, а тому існує ризик неякісної роботи, а під час використання – прийняття рішень, які погіршать транспортну ситуацію в містах.

Також не йдеться про необхідність моніторингу та подальшого керування транспортними потоками в режимі реального часу. А це питання широке і охоплює навігаційні карти, впровадження єдиної цифрової картографічної основи, систему моніторингу транспорту тощо.

Ще одним важливим аспектом впровадження інтелектуальних транспортних систем є підготовка фахівців у цій галузі, які повинні вміти вирішувати задачі оптимізації розподілу транспортних потоків по вулично-дорожній мережі та динамічно спрямовувати маршрути. Також необхідно

оцінити стратегії організації дорожнього руху з урахуванням вимог безпеки руху, екологічних показників та існуючого стану вулично-дорожньої мережі.

З досвіду провідних країн світу можна зробити висновок, що впровадження ІТС потребує системної координації, в якій беруть участь усі органи виконавчої влади та провідні наукові організації. У країнах Західної Європи, США та Японії програми розвитку ІТС приймаються на 5-10 років [4].

В Україні інтелектуальні транспортні системи точково почали використовувати лише в Києві.

Сьогодні найбільш завантажені перехрестя Києва вже аналізує 31 камера з ІТС. Це перехрестя вул. Басейній, вул. Мечникова, бульварі Лесі Українки, Європейська площа та площа Перемоги [2].

За допомогою ІТС у Києві планується зменшити затори та підвищити ефективність роботи громадського транспорту. Інтелектуальна транспортна система аналізує дані з камер і коригує транспортні потоки для максимально ефективного розвантаження вулично-дорожньої мережі та якісної організації громадського транспорту. Планується, що ця система зможе надсилати сповіщення комунальним службам або Патрульній поліції міста Києва для скорочення часу реагування на ДТП та перешкоди на дорозі [2].

Аналіз комплексної схеми руху в Києві проводиться спільно з іноземними експертами. Однією з ключових цілей є оптимізація управління світлофорним об'єктом.

1.2 Аналіз факторів, що впливають на умови дорожнього руху

У фактори, що впливають на швидкість можна узагальнити за двома принципами: внутрішні і зовнішні та статичні і динамічні [4]. Перший класифікується за перспективою транспортного засобу, яка представляє внутрішні фактори транспортного засобу та зовнішні фактори середовища. Останні статичні фактори є постійними, і до них можна отримати доступ онлайн через офлайн-сховище. З іншого боку, динамічні фактори змінюються з часом настільки, що моделювати їх важче, ніж статичні. Таким чином,

історичну та поточну інформацію про динамічні фактори зазвичай слід розглядати разом у прогнозуванні.

До внутрішні фактори належить поведінка водіння (динамічний), інформація про транспортний засіб (статичний), стан автомобіля (динамічний) на рисунку 1.2.

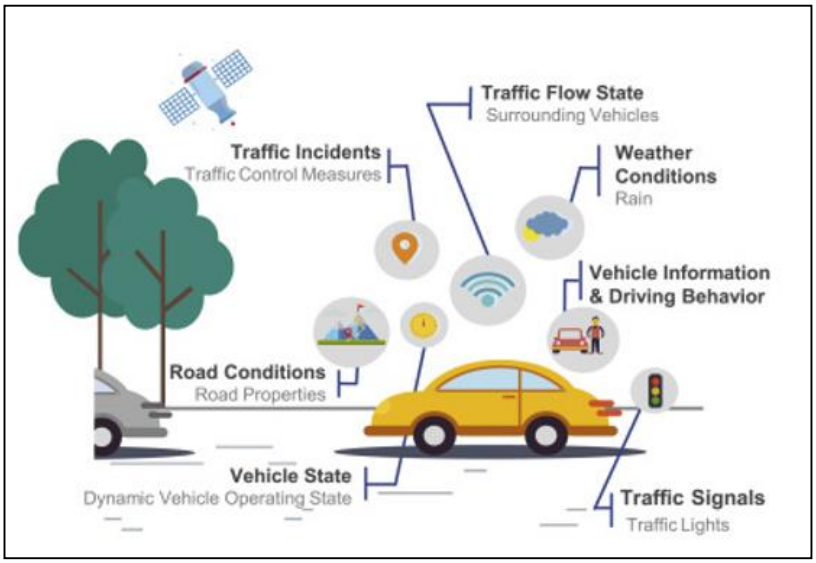

Рисунок 1.2 – Внутрішні фактори впливу на швидкість автомобіля [4]

Водій безпосередньо керує транспортним засобом, і його поведінка значно впливає на швидкість автомобіля. Усі інші фактори впливають на швидкість опосередковано через поведінку за кермом. Однак сприйняття та рішення людини-водія надзвичайно складні. Водію можуть телефонувати, розмовляти та втомлюватися під час водіння. Крім того, в одному сценарії реакції різних водіїв не однакові, і навіть той самий водій може приймати різні рішення.

Інформація про транспортний засіб стосується статичних факторів транспортного засобу, включаючи фізичну інформацію (наприклад, тип транспортного засобу, вага транспортного засобу, вага пасажира та розподіл), силовий агрегат (наприклад, тип потужності, тип трансмісії), потужність (наприклад, максимальне прискорення чи уповільнення, максимальна швидкість, максимальний підйом), стабільність керованості (наприклад,

мінімальний радіус повороту) тощо [4]. Крім того, функції автомобіля також належать до такої інформації, яка контролює швидкість автомобіля відповідно до певних правил.

Наведена вище інформація про транспортний засіб опосередковано обмежує зміну швидкості автомобіля, а динамічні стани автомобіля безпосередньо впливають на швидкість. Ці фактори описують стан транспортного засобу, включаючи швидкість автомобіля, прискорення, наявне паливо або потужність, температуру акумулятора, стан трансмісії тощо.

До зовнішніх факторів належить стан транспортного потоку (динамічний), погодні умови (динамічний), правила дорожнього руху (статичний), світлофори та події (динамічний) на рисунку 1.3.

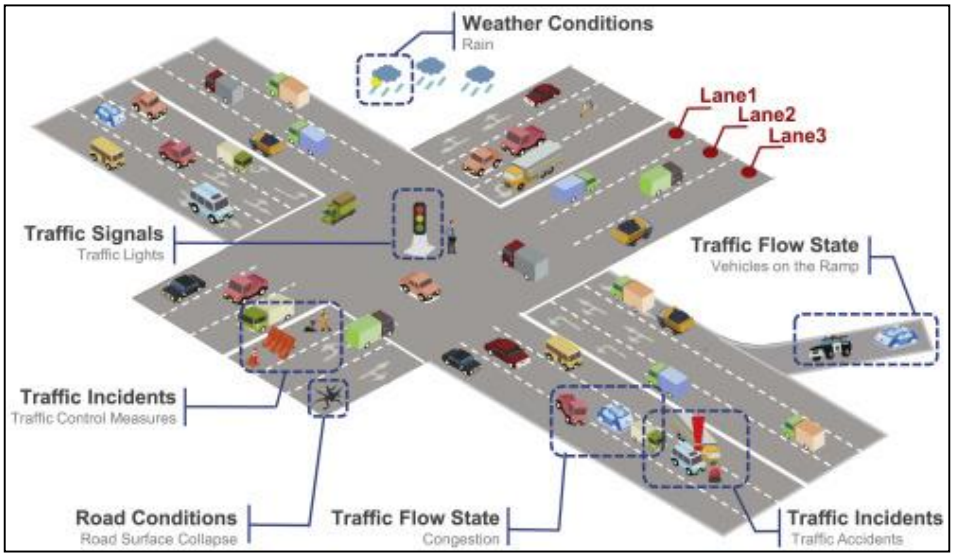

Рисунок 1.3 – Зовнішні фактори впливу на швидкість автомобіля [4]

З мікроскопічної точки зору, стан транспортного потоку стосується стану руху оточуючих транспортних засобів, а стан транспортного засобу, що їде попереду (тобто відносна швидкість і відстань) сильно впливає на швидкість транспортного засобу. З макроскопічної точки зору це стосується стану руху, наприклад, потоку, швидкості та заповненості [4]. Наприклад, затори можуть безпосередньо впливати на швидкість на різних рівнях, і макроскопічні ударні хвилі є фактором впливу. Крім того, стан транспортного

потоку демонструє сильну просторово-часову залежність, і слід враховувати історичний стан транспортного потоку.

Погані погодні умови (наприклад, туман, дощ і сніг) суттєво впливають на видимість і тертя дороги, таким чином змінюючи поведінку водіння і стійкість автомобіля. Наприклад, низька видимість, спричинена сильним туманом або дощем, сильно впливає на сприйняття та контроль транспортних засобів. Таким чином, ці фактори можуть спричинити затори або аварії, які серйозно впливають на швидкість на різних рівнях.

Дороги включають міські дороги, шосе, сільські дороги та деякі функціональні дороги, перехрестя, перехідні ділянки, різкі повороти. Властивості дороги (перепад висоти, кривизна та нерівність), атрибути навколишнього середовища (наприклад, скупчення, школи, лікарні) і правила дорожнього руху (наприклад, нерівність, обмеження швидкості, властивості смуги) також впливають на швидкість.

До дорожніх подій належать дорожньо-транспортні пригоди, заходи з контролю дорожнього руху, соціальні події (наприклад, спортивні заходи, іспити, виступи) тощо. Дорожньо-транспортні пригоди може призвести до заторів, а соціальні події впливають на швидкість, змінюючи попит на трафік. Крім того, світлофори мають важливе значення для підтримки дорожнього руху на міських дорогах, і це серйозно обмежує швидкість у транспортній системі.

Усі перераховані вище фактори впливу впливають на швидкість на різних рівнях, але основні фактори на кожному рівні неоднакові.

Мікрошвидкість відноситься до швидкості транспортного засобу таким чином, що внутрішні фактори щодо станів одного транспортного засобу є основними факторами швидкості транспортного засобу в короткостроковому прогнозуванні, особливо поведінки транспортного засобу.

Що стосується швидкості руху, вплив невизначеності одного транспортного засобу на макрошвидкість послаблюється статистичним усередненням для кількох транспортних засобів. Тому порівняно зі швидкістю

транспортного засобу швидкість руху постійно змінюється і вплив внутрішніх факторів на неї менший. Тим часом, зовнішні чинники стають основними факторами швидкості руху завдяки більшому горизонту прогнозування, ніж у швидкості транспортного засобу.

Більшість досліджень з оцінки транспортного потоку всієї дорожньої мережі ґрунтуються на одній або декількох властивостях дорожньої мережі, і результати можуть не бути багатообіцяючими [5-7], а оцінка ефективності передачі мережі або налаштування параметрів інтелектуальної системи була розглянута в недавніх дослідженнях [8,9].

Існує спосіб поєднання п'яти топологічних показників і довжини дороги для оцінки транспортного потоку на основі підходу множинної регресії [10]. Для оцінки транспортного потоку використовуються шість мір: довжина дороги, близькість, проміжна, ступінь, ранг сторінки та коефіцієнт кластеризації [5].

1.3 Постановка задачі

В результаті аналізу предметної області було визначене основне завдання дослідження. Задача полягає в прогнозуванні інтервалів середньої швидкості автомобілів на певній ділянці дороги на основі історичних даних, тобто даних про швидкість руху автомобілів на цій ділянці за останній часовий період. Метою цієї задачі є надання інформації про те, яка середня швидкість автомобілів очікується на ділянці дороги у певний момент часу в майбутньому.

Тобто, можна виділити чіткі етапи виконання роботи:

- аналіз предметної галузі;
- пошук та аналіз існуючих алгоритмів та методів прогнозування;
- збір даних для дослідження, їх аналіз та передобробка;
- реалізація моделей прогнозування, їх порівняння за обраними метриками;
- висновки, визначення переваг та недоліків кожного алгоритму.

Одним з підходів до розв'язання цієї задачі може бути застосування регресійних моделей, такої як квантильна регресія, для короткострокового прогнозування середньої швидкості автомобілів.

Для цього необхідно мати доступ до історичних даних про швидкість руху автомобілів на цій ділянці дороги або інших важливих факторів, таких як погода, час дня та тижня, додатковий транспортний потік, дорожні роботи та інші. Такий підхід дозволяє отримати точні прогнози з урахуванням можливих відхилень від середніх значень і використовувати їх для покращення систем дорожнього руху та ефективнішого використання дорожньої інфраструктури.

Оцінка точності прогнозу може бути проведена за допомогою різних метрик, таких як середня абсолютна помилка, середня квадратична помилка, середньоквадратична помилка та коефіцієнт детермінації.

Обраний набір даних потрібен бути у вільному доступі та ці дані повинні бути отримані з датчиків, що встановлюються на певній ділянці дороги, і можуть містити інформацію про швидкість автомобілів в різні години доби, дня тижня, що можуть впливати на кількість автомобілів, що рухаються на ділянці дороги.

2 АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ і АЛГОРИТМІВ

2.1 Опис існуючих алгоритмів

Зараз, як правило, існує три методи для реалізації прогнозування потоку трафіку, а саме метод прогнозування параметрів, метод непараметричного прогнозування та метод комбінованого прогнозування параметрів, які показано на рис. 2.1.

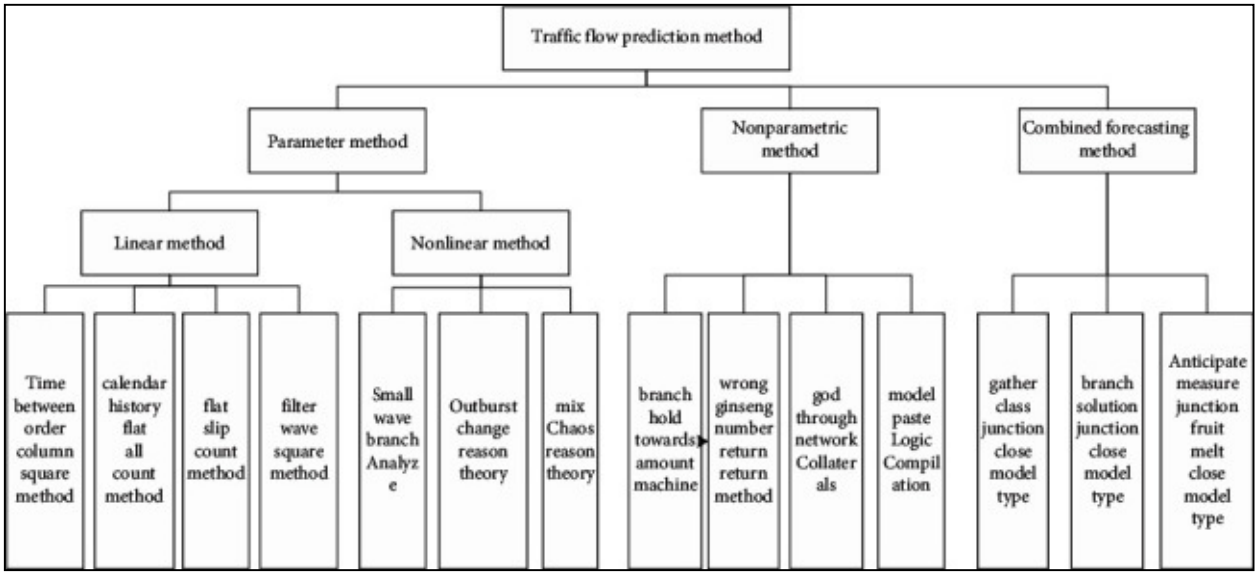

Рисунок 2.1 – Класифікація методів прогнозування транспортного потоку [11]

Методи прогнозування транспортних потоків поділяють на параметричні, непараметричні та комбіновані. Серед них параметричні методи можна розділити на лінійні та нелінійні стилі [11].

Лінійний метод можна підрозділити на метод часових рядів, алгоритм історичного середнього, алгоритм згладжування та алгоритм фільтрації; нелінійні методи можна розділити на вейвлет аналіз, теорію катастроф і теорію хаосу; непараметричні методи включають машину опорних векторів, непараметричну регресію, нейронну мережу та нечітку логіку; комбінований метод прогнозування – поєднання двох або більше методів прогнозування для досягнення ефекту спільного прогнозування. Він застосовує модель комбінації

кластеризації, модель комбінації декомпозиції та технологію моделі об’єднання результатів прогнозування.

Поряд із розробкою структури глибокого навчання на основі нейронної мережі, деякі джерела зосереджені на моделі на основі глибокого навчання, яка використовується для прогнозування потоків трафіку. У [12] автори розглянули моделі глибокого навчання та пов’язані з ними робочі процеси, які використовуються для створення придатної для використання моделі прогнозування. Автори також порівняли моделі глибокого навчання, застосовуючи їх до різних типів завдань прогнозування стану трафіку, використовуючи той самий набір даних, що може дати читачеві більш інтуїтивне уявлення про переваги та недоліки між різними моделями прогнозування глибокого навчання.

Було виявлено, що машинне навчання стає все більш популярним для завдання прогнозування потоку транспорту. Причина полягає в тому, що менше попередніх знань про зв’язок між різними схемами трафіку для побудови моделі, менше обмежень на завдання прогнозування та є кращі нелінійні характеристики.

Існує багато видів методів класифікації моделей машинного навчання на основі різних точок зору [13]:

– регресійна модель: вивчаючи зв’язок між залежною змінною та незалежною змінною, модель регресії намагається використовувати криву або лінію, щоб відповідати набору даних;
– модель на основі прикладу: модель на основі прикладу розв’язує завдання прогнозування шляхом порівняння подібності між вхідною послідовністю та історичними зразками даних і використовує знайдені зразки для створення остаточного прогнозу;
– модель на основі ядра: у моделі на основі ядра ми використовуємо функцію ядра для відображення вхідних даних у векторному просторі високого порядку, де завдання прогнозування легко вирішити;

– модель на основі нейронної мережі: це тип моделі, створений шляхом імітації того, як інформація проходить через нейрони в мозку;
– гібридна модель: у гібридній моделі остаточний прогноз робиться шляхом поєднання двох або більше результатів прогнозування з різних моделей прогнозування.

2.2 Квантильна регресія

Існують різні типи регресії. Регресійні моделі мають на меті підібрати цільову змінну, яка виражається у вигляді числового вектора. Тим не менш, статистики все частіше розробляють складні регресійні техніки регресії. Квантильна регресія (Quantile Regression) – це процедура оцінки параметрів лінійного зв'язку між пояснювальними змінними і заданим рівнем квантиля змінної, що пояснюється [14,15].

На відміну від звичайного методу найменших квадратів, квантильна регресія є непараметричним методом. Це дозволяє отримати більше інформації: параметри регресії для будь-яких квантилів розподілу залежної змінної. Крім того, така модель набагато менш чутлива до викидів у даних і до порушень припущень про характер розподілів.

Нехай Y – це реальна змінна відгуку, а X – коваріативна або предикторна змінна, можливо, високої розмірності. Стандартна мета статистичного аналізу полягає в тому, щоб певним чином визначити зв'язок між Y та X. Стандартний регресійний аналіз намагається отримати оцінку $\hat{\mu}(x)$ умовного середнього значення $E(Y|X = x)$ змінної відгуку Y , якщо $X = x$. Умовне середнє значення мінімізує очікувану квадратичну похибку,

$$E(Y|X = x) = \arg\min E\{(Y - z)^2|X = x\}.$$

Апроксимація умовного середнього зазвичай досягається мінімізацією функції втрат типу квадратичної похибки функції втрат типу квадратичної похибки за наявними даними.

Умовне середнє висвітлює лише один аспект умовного розподілу змінної відгуку Y, але нехтує всіма іншими особливостями що можуть становити інтереса, що призвело до розвитку квантильної регресії. Умовна функція розподілу F(y|X = x) задається формулою ймовірністю того, що для X = x, Y менше, ніж y ∈ R,

$$F(y|X = x) = P(Y \leq y|X = x)$$

Для неперервної функції розподілу α-квантиль $Q_\alpha(x)$ визначається таким чином, що ймовірність того, що Y буде меншою за $Q_\alpha(x)$, для заданого X = x точно дорівнює α. У загальному випадку,

$$Q_\alpha(x) = \inf\{y: F(y|X = x) \geq \alpha\}$$

Квантилі дають більш повну інформацію про розподіл Y як функції предикторної змінної X, ніж лише умовне середнє.

Як приклад, розглянемо прогнози рівня озону на наступний день, як в роботі [10]. Регресія за методом найменших квадратів намагається оцінити умовне середнє значення рівня озону рівня озону. Вона дає мало інформації про коливання рівня озону навколо цього прогнозованого умовного середнього значення. Наприклад, може бути цікаво знайти рівень озону, який з високою ймовірністю не буде перевищений. Цього можна досягти за допомогою квантильної регресії, оскільки вона дає інформацію про розкид змінної відгуку.

Постає питання щодо надійності прогнозу для нового випадку. Знову розглянемо прогноз рівня озону на наступний день. В деякі дні може виявитися можливим визначити рівень озону на наступний день з більшою точністю, ніж в інші дні (це дійсно можна спостерігати для даних про озон, див. розділ з числовими результатами). При стандартному прогнозуванні для кожного нового випадку повертається оцінка в одній точці. Ця оцінка не

містить інформації про дисперсію спостережень навколо прогнозованого значення.

Квантильна регресія може бути використана для побудови інтервалів прогнозування. 95% інтервал прогнозування для значення Y задається формулою:

$$I(x) = [Q_{0.25}(x), Q_{0.975}(x)]$$

Тобто, нове спостереження Y, для X = x, з високою ймовірністю знаходиться в інтервалі I(x). Ширина цього інтервалу прогнозування може сильно змінюватися в залежності від x. Дійсно, повертаючись до попереднього прикладу, можна сказати, що до попереднього прикладу, рівень озону на наступний день можна передбачити в деякі дні в п'ять разів точніше, ніж в інші дні. Цей ефект ще більш виражений для інших наборів даних. Таким чином, квантильна регресія пропонує принциповий спосіб оцінити надійність прогнозів.

Квантильна регресія також може бути використана для виявлення викидів [10]. А нове спостереження (X, Y) вважатиметься викидом, якщо його спостережуване значення Y є екстремальним, в певному сенсі, відносно прогнозованої умовної функції розподілу. Однак не існує загальноприйнятого правила щодо того, що саме є «екстремальним» спостереження. Можна було б позначити спостереження як викиди, якщо відстань між Y та медіаною умовного розподілу є великою, що вимірюється порівняно з якоюсь надійною мірою дисперсії, такою як умовне медіанне абсолютне відхилення або умовного інтерквартильного розмаху. Обидві величини можна отримати за допомогою квантильної регресії.

Таким чином можна виявити лише аномалії в умовному розподілі Y спосіб можна виявити лише аномалії в умовному розподілі Y. Випадкові значення самого X не можуть бути виявлені. Інші дослідження зосереджені на виявленні аномалій для немаркованих даних [10].

Квантильна регресія має на меті оцінити умовні квантилів за даними. Її можна розглядатися як оптимізаційну задачу, так само як оцінка умовного середнього досягається мінімізацією функції втрат квадрата помилки. Нехай функція втрат L_α визначається для $0<\alpha<1$ зваженими абсолютними відхиленнями згідно з формулою (1):

$$L_\alpha(y,\alpha) = \begin{cases} \alpha|y-q| & y \geq q \\ (1-\alpha)|y-q| & y \geq q \end{cases} \quad (1)$$

У той час як умовне середнє мінімізує очікувані квадратичні втрати, умовні квантилі мінімізують очікувані втрати $E(L_\alpha)$,

$$Q_\alpha(x) = \arg\min E\{L_\alpha(Y,q)|X=x\}$$

Параметрична квантильна регресія вирішується шляхом оптимізації параметрів таким чином, щоб емпіричні втрата емпіричних даних була мінімальною.

Цей вид регресії навмисно вносить зміщення в результат замість того, щоб шукати середнє значення прогнозованої змінної, квантильна регресія має на меті знайти медіану та будь-які інші квантилі (які іноді називають процентилями). Класичним і найбільш простим прогнозом є прогноз на основі середніх значень: відповідні ваги завищеного і заниженого прогнозу повинні бути рівними, інакше прогноз стає зміщеним (точніше, зміщеним відносно середнього значення).

Першим уточненням цього підходу є медіанне прогнозування: відповідні частоти надмірного та недостатнього прогнозування повинні бути рівними, інакше прогноз стає зміщеним відносно медіани. На цьому етапі ми змінюємо поняття незміщеного прогнозу з рівних ваг на рівну ймовірність. Цей зсув не є очевидним, але в деяких ситуаціях він може мати величезне числове значення. Медіанне значення являє собою порогове значення, при

якому розподіл розпадається з ймовірністю 50 на 50. Однак можна розглядати й інші співвідношення частот. Наприклад, ми можемо розглядати співвідношення 80 на 20, 90 на 10 і будь-яке інше, якщо їх сумарне значення дорівнює 100%.

Квантилі є узагальненням медіанного значення до будь-якого відсоткового виразу. Для τ, значення якого знаходиться між 0 і 1, квантильна регресія Q(τ) представляє порогове значення, при якому ймовірність значення нижче порогового дорівнює τ [16].

Метод квантильної регресії можна комбінувати або поєднувати з іншими методами для підвищення точності прогнозування, тому в цій статті [17] описано алгоритм короткострокового непараметричного ймовірнісного квантильного регресійного прогнозування, який поєднує переваги гібридної нейронної мережі та квантильної регресії.

Підхід до проблеми квантильної регресії [18,19] з точки зору багатозадачності вирішує неприємну проблему перекриття квантилів, при цьому значно перевершуючи сучасні методи квантильної регресії. У роботах зазначається, що спільне моделювання середнього значення та кількох умовних квантилів призводить до покращення прогнозів умовного математичного сподівання завдяки додатковій інформації та ефектам регуляризації, спричиненим додаванням квантилів.

Також в літературі зустрічаються дослідження з використанням штучних нейронних мереж [20], таких як довга короткочасна пам'ять. Описано стан відсутності даних про швидкість руху та запропоновано метод прогнозування швидкості руху на основі вимірювання транспортного потоку в попередній та наступний моменти часу. Проведено порівняння ефективності п'яти моделей прогнозування: k-найближчих сусідів, регресії опорних векторів (SVR), дерев класифікації, точно довгої короткочасної пам'яті (LSTM) та зворотного поширення (BP) [21]. Метод працює на основі моделі LSTM і досягає найкращого результату.

Загалом, багато досліджень використовують методи квантильної регресії для прогнозування швидкості руху та заторів на дорогах. Використовуються різні методи, такі як метод найближчого сусіда, випадкові ліси, градієнтний бустинг. Кожен з цих методів має свої переваги та недоліки, тому вибір методу залежить від конкретного завдання та обсягу даних, тобто пошуку у сховищах великих даних [22].

2.3 Метод випадкового лісу

Випадковий ліс – це широко використовуваний алгоритм класифікації та регресії. Оскільки класифікація і регресія є найбільш важливими аспектами машинного навчання, можна сказати, що алгоритм випадкового лісу є одним з найважливіших алгоритмів машинного навчання.

Дерево регресії є непараметричним методом прогнозування, який наближає умовне середнє значення, використовуючи наявні дані, що знаходяться в безпосередній близькості до точки, яку ми хочемо передбачити. Для неперервних предикторів дерева регресії розбивають простір предикторів на прямокутники високої розмірності, а не використовують сусідів.

Випадкові ліси вирощують ансамбль дерев, використовуючи n незалежних спостережень (Y_i, X_i), i=1,…n. Вирощується велика кількість дерев. Для кожного дерева і кожного вузла випадкові ліси використовують випадковість при виборі змінної для розбиття. Для кожного дерева використовується пакетна версія навчальних даних. Крім того, тільки випадкова підмножина предикторних змінних розглядається для вибору точки розщеплення в кожному вузлі. Розмір випадкової підмножини, який називається mtry, є єдиним параметром налаштування алгоритму, хоча результати, як правило, майже оптимальні в широкому діапазоні цього параметра. Значення mtry можна точно налаштувати на вибірках поза пакетом [10].

Прогноз випадкових лісів для нової точки даних X = x є усередненою відповіддю всіх дерев. Алгоритм дещо схожий на пов'язаний з бустінгом, де дерева виступають у ролі учнів. Однак, у випадку з випадковими лісами кожне

дерево вирощується на основі початкових спостережень змінної відгуку, тоді як бустінг намагається підігнати залишки з урахуванням прогнозу раніше згенерованих дерев.

Дотримуючись умовних позначень, назвемо θ вектором випадкових параметрів, який визначає, як вирощується дерево (наприклад, які змінні враховуються для точок розщеплення у кожній вершині). Відповідне дерево позначимо через T(θ). Прогноз окремого дерева T(θ) для нової точки даних X = x отримується шляхом усереднення спостережуваних значень у листі l(x, θ). Нехай вектор ваги w_i(x, θ) задано додатною константою, якщо спостереження X_i є частиною листа l(x, θ), і 0, якщо ні. Ваги дорівнюють одиниці згідно з формулою (2):

$$w_i(x, \theta) = \frac{1_{\{X_i \in R_{l(x,\theta)}\}}}{\#\{j : X_j \in R_{l(x,\theta)}\}} \quad (2)$$

Використовуючи випадкові ліси, умовне середнє значення E(Y | X = x) апроксимується усередненим прогнозом k окремих дерев, кожне з яких побудовано за допомогою вектора θ_t , t = 1 ..., k. Нехай w_i(x) є середнім значенням w_i(θ) для цього набору дерев згідно з формулою (3):

$$w_i(x) = k^{-1} \sum_{t=1}^{k} w_i(x, \theta_t) \quad (3)$$

Тож, прогнозування випадкових лісів має такий вигляд:

$$\hat{\mu}(x) = \sum_{t=1}^{k} w_i(x) Y_i$$

Наближення умовного середнього значення Y при X = x, таким чином, є зваженою сумою всіх спостережень. Ваги залежать від коваріати X = x і мають тенденцію бути великими для тих i ∈ {1,..., n}, де умовний розподіл Y, заданий X = X_i, подібний до умовному розподілу Y при X = x [10].

Модель випадкового лісу (Random Forest) є дуже цінною і прикладною непараметричною формою регресії. Дерева забезпечують природний спосіб автоматичної апроксимації f(X) без особливих роздумів про те, як виглядає істинна функція f(X). Його мішкова природа забезпечує кращу точність прогнозування, ніж дерево регресії, а також дозволяє включати категоричні предиктори там, де інші непараметричні інструменти, цього не роблять.

2.4 Метод k-найближчих сусідів

K-найближчих сусідів (KNN) – це інструмент непараметричної регресії, який намагається оцінити умовне середнє для нового спостереження x_0 шляхом визначення k точок спостережуваних даних, які є найближчими до нового спостереження, для якого потрібен прогноз. Потім значення відгуку для цих найближчих спостережень усереднюються разом [23-26].

Прогнози k найближчих сусідів більш формально обчислюються за допомогою наступного рівняння:

$$\hat{Y} = \frac{1}{k}\sum_{x_i \in N_k(x_0)} y_i,$$

де $N_k(x_0)$ – околиця x_0, визначена k найближчими точками x_i у навчальних даних.

Оскільки більшість спостережуваних даних, ймовірно, матимуть лише одне спостереження або не матимуть жодного спостереження в точці x_0, спостережувані значення відгуку для найближчих сусідів слугують наближеним значенням умовного розподілу $Y|x = x_0$. Таким чином, усереднення за цими спостережуваними значеннями є оцінкою умовного середнього значення в точці x_0 [16].

Застосування алгоритму KNN у короткостроковому прогнозуванні міського трафіку. Алгоритм KNN має хороші показники при роботі з

раптовими змінами та нелінійністю міського транспортного потоку завдяки своїм непараметричним регресійним характеристикам [27].

Однак тривалий час виконання системи прогнозування KNN призводить до зниження ефективності прогнозування. Експериментальні результати показують, що цей метод ефективно покращує ефективність прогнозування системи за умови гарантування точності вихідного прогнозу. Ідеї, представлені тут, можуть бути далі досліджені з додатковими даними, такими як погодні умови або надзвичайні ситуації, більш складні міські топології та різні типи методів прогнозування .

2.5 Метод Gradient Boosting Machines

Gradient Boosting Machines (GBM) – це ансамблі деревовидних методів, які використовують принцип підсилення слабких учнів (найчастіше алгоритм бінарного дерева рішень) з використанням архітектури градієнтного спуску [28]. Зокрема, на кожній ітерації нова модель навчається так, щоб мінімізувати помилку ансамблю, навченого до цього моменту. У свою чергу, XGBoost є вдосконаленням фреймворку GBM шляхом оптимізації системи та вдосконалення алгоритму.

Найважливішим параметром градієнтного бустингу є кількість ітерацій, що виконуються для підбору алгоритму. Занадто мала кількість ітерацій може призвести до неоптимального підбору, а занадто велика – до надмірного.

Хоча існують різні підходи до оптимізації параметрів алгоритму, ці підходи вимагають значних обчислювальних витрат на великих масивах даних. Іншими недоліками градієнтного бустингу є, те що вони вимагають багато пам'яті через велику кількість ітераційі вони повільніше навчаються порівняно з іншими методами.

У роботах [29,30] показано використання квантильної регресії для прогнозування трафіку на основі даних зі смартфонів. Вони порівняли різні методи квантильної регресії, включаючи найближчих сусідів, випадкові ліси та градієнтний бустинг, і виявили, що метод градієнтного бустингу дає найкращі результати. А також ряд статистичних методів для прогнозування 5-

го, 10-го, 25-го, 50-го, 75-го і 90-го процентилів швидкості руху. В результаті порівняння моделей автори дійшли висновку, що метод найближчого сусіда та випадкові ліси показали найкращі результати для прогнозування трафіку з використанням квантильної регресії.

2.6 Алгоритм вирахування квантилів

У цій роботи пропонується інший підхід, який безпосередньо не використовує мінімізацію функції втрат виду (1). Алгоритм описано на основі методу випадкового лісу, але він підходить і для інших методів квантильної регресії.

Випадкові ліси вирощують ансамбль дерев, використовуючи випадковий вибір вузлів і точок розщеплення. Прогнозування випадкових лісів можна розглядати як адаптивну процедуру класифікації та регресії околиць. Для кожного X = x, набір ваг $w_i(x)$, i = 1,..., n для початкових n спостережень. Прогнозування випадкових лісів або оцінка умовного середнього є еквівалентна середньозваженому значенню спостережуваних змінних відгуку. Для квантильних регресійних лісів дерева вирощуються так само, як і в стандартному алгоритмі випадкових лісів. Потім умовний розподіл оцінюється за допомогою зваженого розподілу спостережуваних змінних відгуку, де ваги, привласнені спостереженням, ідентичні оригінальному алгоритму випадкових лісів.

Вище було показано, що випадкові ліси апроксимують умовне середнє E(Y |X = x) середньозваженим значенням за спостереженнями змінної відгуку Y. Можна було б припустити, що зважені спостереження дають не тільки хороше наближення до умовного середнього, але й до повного умовного розподілу середнього, але й до повного умовного розподілу. Функція умовного розподілу Y, за умови, що X = x, має вигляд:

$$F(y|X = x) = P(Y \leq y|X = x) = E(1_{\{Y \leq y\}}|X = x)$$

Останній вираз підходить для проведення аналогії з апроксимацією випадковим лісом умовного середнього E(Y |X = x). Подібно до того, як E(Y |X = x) апроксимується середньозваженим значенням за спостереженнями Y, визначимо наближення до $E(1_{\{Y\leq y\}}|X=x)$ середньозваженим за спостереженнями $1_{\{Y\leq y\}}$ за формулою (4):

$$\hat{F}\,(y|X=x)=\sum_{i=1}^{n} w_i(x)1_{\{Y\leq y\}} \qquad (4)$$

З використанням тих самих ваг $w_i(x)$, що і для випадкових лісів, визначених у рівнянні (3). Це наближення лежить в основі алгоритму квантильних регресійних лісів.

Алгоритм обчислення оцінки $\hat{F}(y|X = x)$ можна узагальнити так:

а) побудувати k дерев $T(\theta_t)$, t = 1,…,k, як у випадкових лісах, але для кожного листка кожного дерева треба занотувати всі спостереження у цьому листі, а не лише їх середнє значення;

б) для заданого значення X = x скинути x з усіх дерев. Обчислити вагу $w(x, \theta_t)$ спостереження $i \in \{1,\ldots, n\}$ для кожного дерева, як у (2). Обчислити вагу $w_i(x)$ для кожного спостереження $i \in \{1,...,n\}$ як середнє по $w_i(x, \theta_t)$, t = 1,...,k, як у (3);

в) обчислити оцінку функції розподілу, як у (4), для всіх $y \in R$, використовуючи вагами з кроку б.

Оцінки $\hat{Q}_\alpha(x)$ умовних квантилів $Q_\alpha(x)$ отримуємо, підставляючи $\hat{F}$ (y|X = x) замість F(y|X = x) у (1). Ключова відмінність між квантильними лісами регресії та випадковими лісами полягає в наступному: для кожного вузла в кожному дереві випадкові ліси зберігають лише середнє значення спостережень, які потрапляють в цей вузол, і нехтують всією іншою інформацією. На відміну від цього, квантильні регресійні ліси зберігають значення всіх спостережень у цьому вузлі, а не лише їхнє середнє, і оцінюють умовний розподіл на основі цієї інформації.

3 ПЛАНУВАННЯ ЕКСПЕРИМЕНТАЛЬНОЇ ЧАСТИНИ ДОСЛІДЖЕННЯ

3.1 Опис набору даних

З наявних даних можна зробити висновок, що набори даних з інформацією про середню швидкість автомобілів використовуються для моніторингу та аналізу моделей руху в різних місцях. Ці набори даних зазвичай збираються за допомогою різних засобів, таких як датчики, камери та мобільні пристрої, і використовуються транспортними організаціями, дослідниками та іншими зацікавленими сторонами для оптимізації транспортних потоків та зменшення заторів.

Однією з головних проблем у використанні наборів даних зі швидкісною інформацією є обмежена доступність відкритих наборів даних. Хоча існує багато джерел даних про дорожній рух, включаючи приватні компанії та державні установи, доступ до цих даних часто обмежений через конфіденційність або право власності.

У багатьох випадках набори даних зі швидкістю можуть збиратися приватними компаніями, які спеціалізуються на моніторингу та аналізі дорожнього руху. Ці компанії можуть неохоче надавати свої дані безкоштовно, оскільки вони покладаються на ці дані як на ключовий компонент своєї бізнес-моделі. Аналогічно, державні органи можуть мати побоювання щодо оприлюднення конфіденційних даних, які можуть поставити під загрозу приватне життя людей або розкрити стратегічну інформацію про транспортну інфраструктуру.

Оскільки для побудови моделі було вирішено використовувати дані по Києву, було вирішено шукати необхідну інформацію на ресурсах відомих служб таксі. У Києві є кілька великих служб таксі, одна з найбільших – Uber [31].

Uber Movement – це сервіс, який надає доступ до відкритих даних про дорожній рух і покликаний допомогти транспортним органам, містобудівникам та дослідникам приймати обґрунтовані рішення щодо

планування та управління транспортом. Сервіс використовує анонімні дані з мільйонів поїздок Uber, щоб дати уявлення про схеми руху та час у дорозі в містах по всьому світу. Був запущений у 2018 році і містить дані з січня 2018 року по березень 2020 року (див. рис. 3.1).

Однією з переваг Uber Movement є те, що він надає доступ до великого і постійно поновлюваного набору даних, які отримуються з фактичних поїздок автомобілів. Це забезпечує рівень деталізації та точності, якого важко досягти за допомогою інших джерел даних, таких як дорожні датчики або камери. Дані розбиті на набори, кожен з яких містить інформацію про середню швидкість руху таксі на відрізку дороги регіону за кожну годину кожного дня конкретного місяця. Для включення даних до моделі вони мають містити інформацію про щонайменше 5 унікальних поїздок на досліджуваній ділянці у відповідний момент часу. Хоча дані, що надаються Uber Movement, мають певні обмеження, наприклад, вони відображають лише частину загального трафіку в певному регіоні.

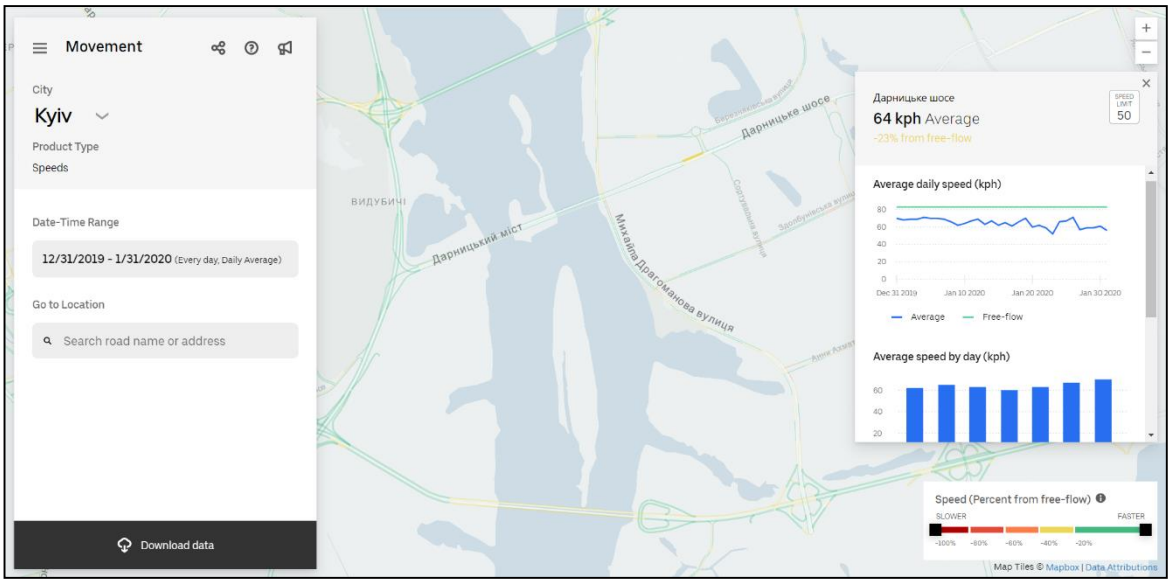

Рисунок 3.1 – Інструмент веб-дослідження швидкості руху Uber для Києва [31]

Ще однією перевагою Uber Movement є те, що дані надаються у стандартизованому форматі, що дозволяє легко порівнювати та аналізувати

дані з різних джерел. Це може бути особливо цінним для дослідників і транспортних органів, яким потрібно проаналізувати дані з кількох міст або регіонів. Набір даних включає наступні поля:

- рік – рік спостереження;
- місяць – номер місяця спостереження (від 1, що відповідає січню, до 12, що відповідає грудню);
- день – день спостереження (від 1 до 31);
- година – година спостереження за місцевим часом (від 0 до 23);
- utc_timestamp – дата та час спостереження у форматі UTC (всесвітній координований час);
- osm_way_id – ідентифікатор дороги OpenStreetMap для відповідного сегмента;
- osm_start_node_id – відповідний ідентифікатор вузла OpenStreetMap для початку сегмента;
- osm_end_node_id – відповідний ідентифікатор вузла OpenStreetMap для кінця сегмента;
- speed_kph_mean – середня швидкість автомобілів Uber на відповідній ділянці дороги у км/год;
- speed_kph_stddev – середньоквадратичне відхилення швидкості на відповідній ділянці дороги у км/год.

3.2 Вибір методів дослідження

Хоча існують багато методів для прогнозування трафіку, кожен з них має свої переваги та недоліки. Вибір квантильної регресії для прогнозування дорожнього трафіку є доцільним з кількох причин.

По-перше, вона дозволяє оцінювати не тільки середнє значення залежної змінної, а й кілька квантилів розподілу цієї змінної. Це важливо для прогнозування дорожнього трафіку, оскільки швидкість руху на дорозі може бути досить різною в різних точках та часах. Наприклад, у пік години трафік може бути дуже густим, що призводить до збільшення часу, необхідного для проїзду відрізка дороги. Квантильна регресія дозволяє моделювати різні

квантилі розподілу часу проїзду в залежності від різних факторів, таких як день тижня, час дня, погода тощо.

По-друге, квантильна регресія дозволяє моделювати залежність між залежною та незалежною змінними з урахуванням нелінійності та взаємодії між факторами. Це важливо для прогнозування дорожнього трафіку, оскільки швидкість руху на дорозі може залежати від багатьох факторів, таких як кількість автомобілів на дорозі, наявність світлофорів, дорожніх знаків тощо.

По-третє, квантильна регресія є менш чутливою до викидів в даних, ніж класична лінійна регресія, оскільки квантильні оцінки є менш чутливими до значень, які відхиляються від середнього значення. Це важливо для прогнозування дорожнього трафіку, оскільки воно дозволяє врахувати невизначеність та випадковість даних про трафік. Такі методи можуть бути особливо корисними для прогнозування піків навантаження, коли зміна одного фактора (наприклад, погодні умови) може призвести до значного збільшення або зменшення трафіку.

Виходячи з обраного набору даних було обрано наступні методи машинного навчання для їх подальшого аналізу: k-найближчих сусідів, випадкового лісу та градієнтного бустингу.

Кожен з цих методів має свої переваги та недоліки, які можуть впливати на його ефективність при прогнозуванні дорожнього трафіку. Наприклад, KNN добре працює на невеликих наборах даних, але може стати повільним на великих масштабах. Випадковий ліс зазвичай дає досить точні прогнози, але може бути менш ефективним на наборах даних з більшою кількістю факторів. Градієнтний бустинг може працювати краще, ніж випадковий ліс на великих наборах даних з багатьма факторами, але може вимагати більшої обчислювальної потужності.

Загалом, вибір методу для прогнозування дорожнього трафіку залежить від конкретних умов і вимог задачі. Потрібно ретельно проаналізувати набір даних та умови задачі для того, щоб вибрати оптимальний метод. Часто

використовують комбінації з декількох методів, які доповнюють один одного, для отримання кращих результатів.

3.3 Метрики оцінювання

Оцінка точності прогнозу може бути проведена за допомогою різних метрик, таких як середня абсолютна відсоткова помилка (MAPE), середня абсолютна помилка (MAE), середньоквадратична помилка (RMSE), коефіцієнт детермінації (R^2). Вони розраховуються наступним чином:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|,$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i},$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

де N – кількість точок даних, y_i – спостережене значення, $\hat{y}_i$ – прогнозоване значення.

4 ПРОВЕДЕННЯ ЕКСПЕРИМЕНТАЛЬНОЇ ЧАСТИНИ ДОСЛІДЖЕННЯ

4.1 Аналіз та підготовка набору даних

Загалом, перший крок – це збір даних. Це можуть бути дані про деякі характеристики руху, що можна зробити за допомогою спеціальних пристроїв, які можуть бути встановлені як зовні транспортного засобу, наприклад, радари, так і всередині.

Системи розпізнавання часто використовуються для додавання даних з камер спостереження для визначення трафіку на дорогах за допомогою технології комп'ютерного зору. Дані можуть бути зібрані в одній точці або на частині дороги, або на множині точок спостереження чи ділянок дороги.

Для проведення експерименту було використано сервіс Uber Movement, який надає дані у відкритому доступі, зокрема для академічних цілей [31]. Тут постає один неочевидний аспект, з одного боку, швидкість таксі, як правило, відображає швидкість руху транспортного потоку, в якому рухається автомобіль. Однак іноді це не так, зокрема, існує дослідження, яке показує, що таксі рухаються дещо повільніше, ніж транспортний потік, з чого логічно випливає, що таксі сповільнюють рух транспорту.

Для проведення експерименту було завантажено набір даних Uber Movement Speed Data з Києва за січень та лютий 2020 року та відповідний .geojson файл з даними OpenStreetMap.

Також було проаналізовано вплив набору даних на оцінку прогнозування міського руху, і експериментальні результати показують, що прогностичний ефект багатомасштабної моделі набагато кращий, ніж у одномасштабного прогнозування, і повністю відображає набір даних, додавання більшої кількості інформації має більшу дослідницьку цінність. Ці ресурси можуть забезпечити подальше розуміння та перспективи використання методів науки про дані та машинного навчання для прогнозування та аналізу транспортних моделей і тенденцій.

Для дослідження були обрані дані головних вулиць центральної частини Києва. У наборі з метаінформацією про сегмент важливість вулиці визначається параметром osmhighway.

Зокрема, значення trunk, primary та secondary позначають головні дороги, а trunk_link, primary_link та secondary_link – основні зв'язки між вулицями, які не мають власної назви, наприклад, з'їзди з естакад або шляхопроводів.

Дороги, які відповідають цьому опису, показані на рисунку 4.1. Це магістральні вулиці, обмежені з одного боку так званою «малою кільцевою дорогою», а з іншого – річкою Дніпро.

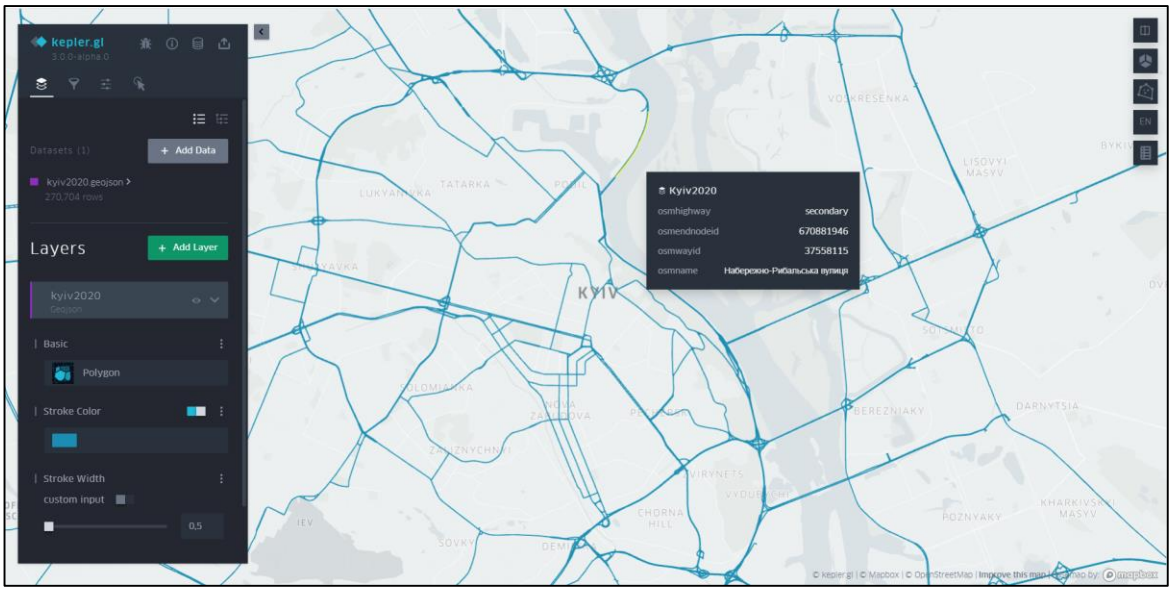

Рисунок 4.1 – Відрізки головних вулиць у центральній частині Києва [32].

Основним інструментом програмування в цій роботі є мова Python з кількох причин. По-перше, вона досить проста у використанні, оскільки має інтуїтивно зрозумілий синтаксис. Таким чином, вона широко використовується професіоналами на всіх рівнях їхньої наукової/інженерної кар'єри. По-друге, вона має великий вибір бібліотек та фреймворків для моделей машинного навчання. І остання причина полягає в тому, що це сучасна мова програмування, яка може бути легко інтегрована з іншими мовами, якщо це необхідно. Для моделювання алгоритму прогнозування було

використано Python 3.10 та Jupyter Notebook. Бібліотеки для аналізу та візуалізації даних – NumPy, Pandas, Sklearn, matplotlib, seaborn.

Для коректної роботи необхідно попередньо обробити досліджувані дані, зокрема, якимось чином відновити відсутні значення. Це може бути пов'язано зі помилками датчиків або, наприклад, відсутністю автомобілів на заданій ділянці дороги в заданий час у випадку з даними служб таксі.

По-друге, дослідникам також важливо звертати увагу на те, як організовані дані про рух на дорожній мережі. З огляду на своє походження, вони зазвичай містять часові та просторові залежності різної складності. Зокрема, для них характерна сезонність у часі, наприклад, як добова, так і тижнева сезонність. Рисунок 4.2 показує характерну сезонність для одного сегменту дороги, що повторюється і на інших.

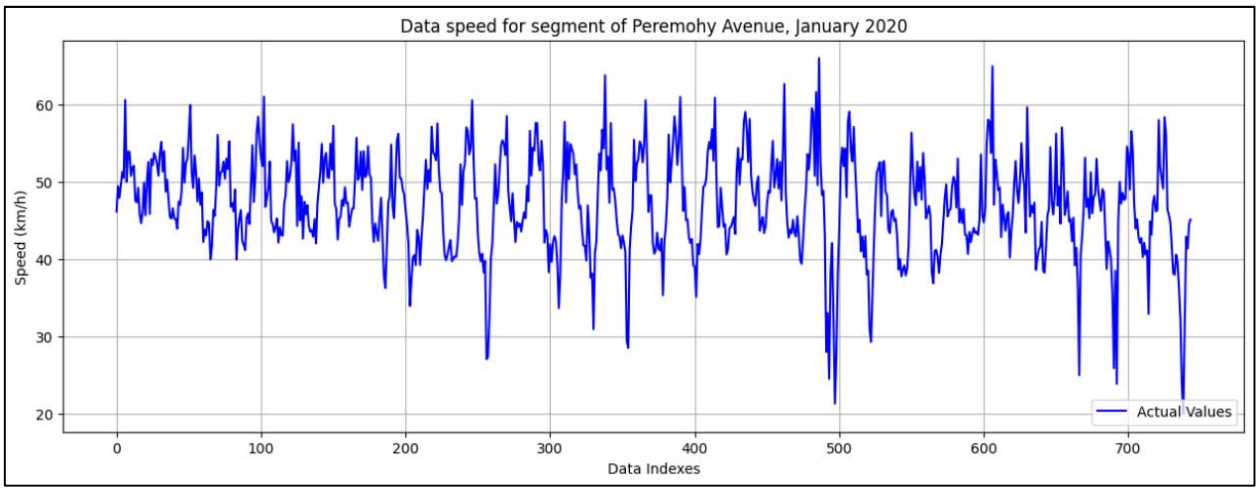

Рисунок 4.2 – Дані з сегменту проспекта Перемоги за січень 2020 року (Рисунок виконаний самостійно)

Можна побачити добову сезонність, тобто те що вдень середня швидкість автомобілів зменшується, а вночі відповідно збільшується (рис. 4.3).

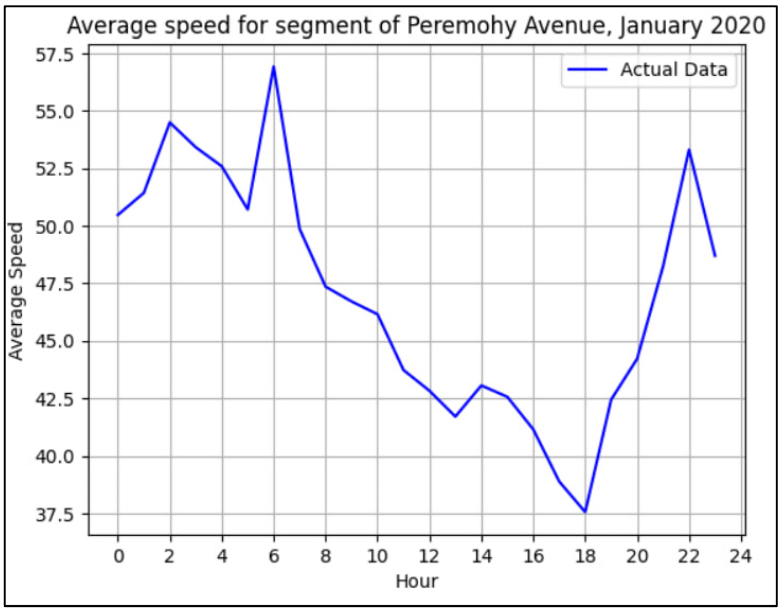

Рисунок 4.3 – Дані з сегменту проспекта Перемоги за січень 2020 року (Рисунок виконаний самостійно)

Також можна побачити залежність швидкості автомобіля з днями тижня (рис. 4.4). Так у вихідні дні, а саме у неділю найшвидший рух на цьому сегменті дороги в порівняння з іншими днями.

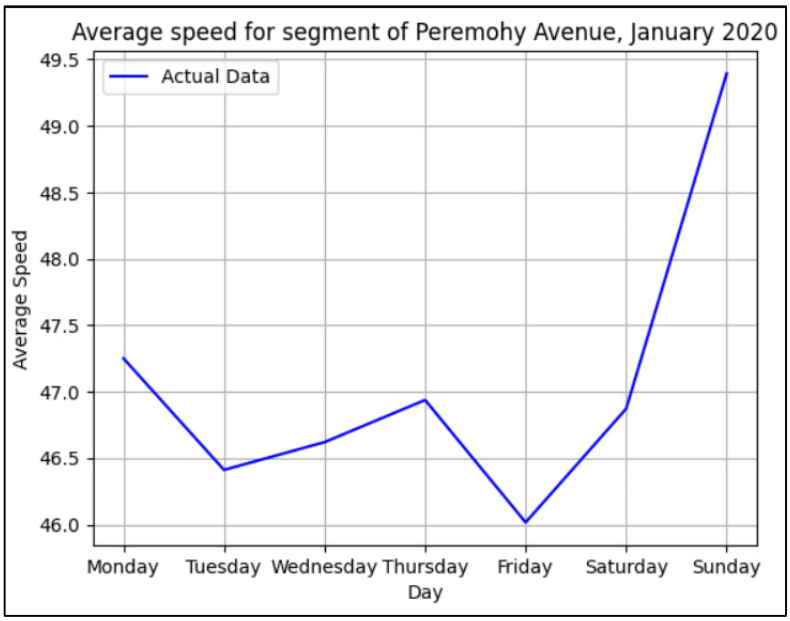

Рисунок 4.4 – Дані з сегменту проспекта Перемоги за січень 2020 року (Рисунок виконаний самостійно)

Сусідні ділянки доріг також впливають на трафік цільового сегмента, що свідчить про очевидні просторові залежності в такого роду даних. Якщо ігнорувати ці аспекти, важко побудувати адекватну систему прогнозування. Тому було видалено зайві поля, такі як рік та місяць, оскільки набір даних містить дані лише за один рік та місяць, а також ідентифікатор відрізків дороги, початковий та кінцевий відрізки, оскільки вони дублювалися.

Зібрані дані часто не є досконалими і не можуть бути використані одразу. Вони можуть містити прогалини або непотрібну інформацію, яка перевантажить алгоритм, і в результаті алгоритм не дасть точного прогнозу. Прогалини в зібраних телекомунікаційних даних виникають з різних причин: системні проблеми, втрата пакетів, перешкоди тощо. Інші проблеми, пов'язані з даними, включають помилки вимірювання датчиків, викиди та прогалини.

4.2 Опис експерименту

Для експерименту були взяті дані з проспекта Перемоги в місті Київ. Вулиця довжиною в 11.8 км, другий найдовший проспект міста, характеризується досить високою транспортною завантаженістю, оскільки є однією з головних вулиць. Набір даних містять 62 сегменти вулиці, де в середньому 690 спостережень за місяць та від 14 до 24 вимірювань за день. Даний набір даних містить інформацію тільки погодинно, що базується на 5 унікальних поїздок на досліджуваній ділянці у відповідний момент часу. Тож обраними цільовими змінними для роботи обрано характеристики – час (година), день (доба). В якості квантилів було обрано 0.1, 0.5 та 0.9.

Для будування прогнозу використовувалися історичні дані, що відповідають 8 тижням року, тобто з 1 січня по 25 лютого 2020 року. Усі дані були розділені на навчальну та тестову вибірки, що відповідають 6 та 2 тижням відповідно. Для кожної моделі будувався прогноз на добу.

Були реалізовані моделі для 3 алгоритмів KNN, випадковий ліс та градієнтний бустинг відповідно. Налаштування параметрів усіх моделей експерименту наведено в таблиці 4.1.

Таблиця 4.1 – Опис параметрів алгоритмів (Таблиця виконана самостійно)

Назва алгоритму	Використані параметри
K-nearest neighbors	neighbors = 5, weights='uniform'
Random Forest	n_estimartors = 100, criterion='squared_error', min_samples_split=2, min_samples_leaf=1, max_depth = 10
Gradient Boosting	n_estimartors = 100, max_depth = 10, random_state = 42, loss=’quantile’, alpha=q, min_samples_leaf=9, learning_rate=0.01

Також було визначено функції для обчислення квантильних втрат і побудови графіків прогнозованих і фактичних значень. Якість прогнозування можна оцінити за допомогою показників MAE, MAPE, RMSE та R^2 з бібліотеки sklearn.metrics.

Машинне навчання вимагає багато оперативної пам'яті. Щоб прискорити доступ до неї, потрібен процесор, який підтримує чотири канали, а не два, як у звичайних кастомних рішеннях. Для ефективного виконання машинного навчання важливо враховувати кількість ядер і обсяг пам'яті відеокарти.

Час тренування на стіні моделі сильно залежить від обчислювальної потужності, був взятий з процесором Intel Core i7 і 8 ГБ оперативної пам'яті, який використовувався для тестування часу тренування на стіні з заданими параметрами.

4.4 Отримані результати

В результаті експерименту було отримано довірчий інтервал – 90% для 62 сегментів дороги. Для прикладу отриманих даних показано наступний графік фактичних та прогнозованих даних прогнозу на добу, а саме на 15 лютого за допомогою алгоритму випадкового лісу (рис. 4.5).

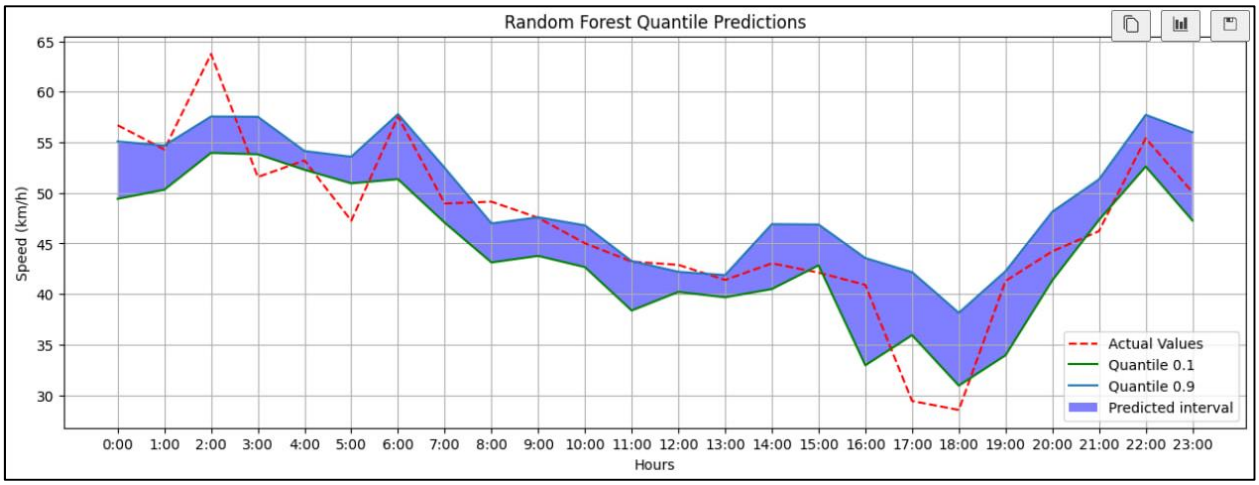

Рисунок 4.5 – Графік квантильної регресії випадкового лісу з прогнозованими значеннями на добу (Рисунок виконаний самостійно)

У таблиці 4.2 наведено більш детальну інформацію про прогнозовані значення за допомогою регресії випадкового лісу для всіх квантилів.

Таблиця 4.2 – Фрагмент прогнозованих значення за допомогою методу випадкового лісу (Таблиця виконана самостійно)

Фактичне значення	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
56.694	49.413	54.718	55.088
54.296	50.312	50.899	54.681
63.731	53.956	54.858	57.557
51.569	53.82	54.508	5753
53.206	52.279	53.594	54.139
47.248	50.941	51.235	53.575
57.56	51.352	53.574	57.796
48.95	47.048	48.681	52.456

Вміст таблиці 4.3 показує метрики помилки для прогнозованих значень за допомогою методу випадкового лісу для одного сегменту дороги для всіх квантилів.

Таблиця 4.3 – Метрики помилки для прогнозованих значень за допомогою методу випадкового лісу (Таблиця виконана самостійно)

Метрика помилки	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
R^2	0.6627	0.8268	0.6578
MAE	3.9356	2.4912	3.4259
MAPE	0.0859	0.0571	0.0837
RMSE	21.3805	10.9803	21.6949

Усереднені метрики помилки можуть також допомогти в порівнянні ефективності різних моделей, що використовуються для прогнозування на одних і тих же даних. Наприклад, якщо одна модель має значно меншу середню квадратичну помилку, ніж інша модель, то можна припустити, що перша модель прогнозує точніше на різних діапазонах значень. Вміст таблиці 4.4 показує усереднені метрики помилки для прогнозованих значень за допомогою методу випадкового лісу для всіх сегментів.

Таблиця 4.4 – Метрики помилки для прогнозованих значень за допомогою методу випадкового лісу (Таблиця виконана самостійно)

Метрика помилки	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
R^2	0.1345	0.6985	0.401
MAE	4.1335	3.2073	5.1086
MAPE	0.0871	0.0885	0.147
RMSE	47.8818	33.5862	66.5557

Час прогнозування для цієї моделі склав 1.205 с., а загалом для всіх 62 моделей 30.228с. Тобто в середньому прогнозування однієї моделі зайняло 0. с.

Потім було отримано наступний графік фактичних і прогнозованих даних на добу, а саме на 15 лютого за допомогою алгоритму k-найближчих сусідів (див. рис. 4.6).

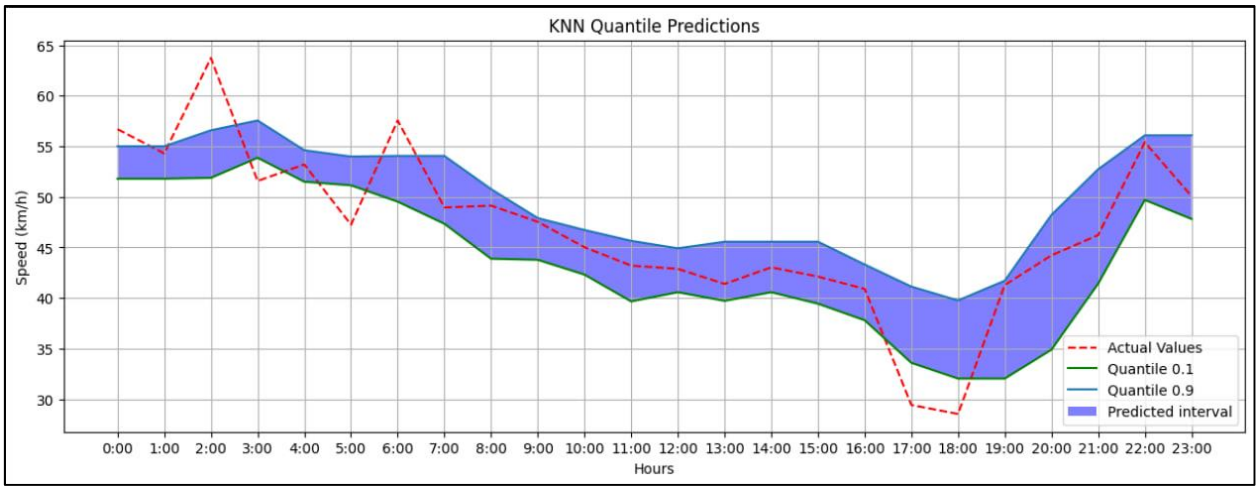

Рисунок 4.6 – Графік квантильної регресії KNN з прогнозованими значеннями на добу (Рисунок виконаний самостійно)

Вміст таблиці 4.5 наведено більш детальну інформацію про прогнозовані значення за допомогою регресії k-найближчих сусідів для всіх квантилів.

Таблиця 4.5 – Фрагмент прогнозованих значення за допомогою методу k-найближчих сусідів (Таблиця виконана самостійно)

Фактичне значення	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
56.694	51.793	54.376	55
54.296	51.793	54.224	55
63.731	51.877	54.867	56.579
51.569	53.868	54.867	57.556
53.206	51.493	53.63	54.61
47.248	51.14	52.218	53.986

Кінець таблиці 4.5

Фактичне значення	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
57.56	49.527	52.218	54.05
48.95	47.337	52.218	54.05

Вміст таблиці 4.6 показує метрики помилки для прогнозованих значень за допомогою методу k-найближчих сусідів для одного сегменту дороги для всіх квантилів.

Таблиця 4.6 – Метрики помилки для прогнозованих значень за допомогою методу k-найближчих сусідів (Таблиця виконана самостійно)

Метрика помилки	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
R^2	0.5944	0.7518	0.6111
MAE	4.3057	2.9189	3.9002
MAPE	0.0926	0.0689	0.0939
RMSE	25.712	15.7311	24.6562

Вміст таблиці 4.7 показує усереднені метрики помилки для прогнозованих значень за допомогою методу k-найближчих сусідів для всіх сегментів.

Таблиця 4.7 – Метрики помилки для прогнозованих значень за допомогою методу k-найближчих сусідів (Таблиця виконана самостійно)

Метрика помилки	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
R^2	0.1987	0.754	0.2929
MAE	4.6294	3.1664	5.8065

Кінець таблиці 4.7

Метрика помилки	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
MAPE	0.0906	0.0866	0.1808
RMSE	47.4401	30.452	86.3454

Час прогнозування для цієї моделі склав 0.001 с., а загалом для всіх 62 моделей зайняло 0.62 с. Тобто в середньому прогнозування однієї моделі зайняло 0.009 с.

Потім було отримано наступний графік фактичних і прогнозованих даних на добу, а саме на 15 лютого за допомогою алгоритму градієнтного бустингу (див. рис. 4.7).

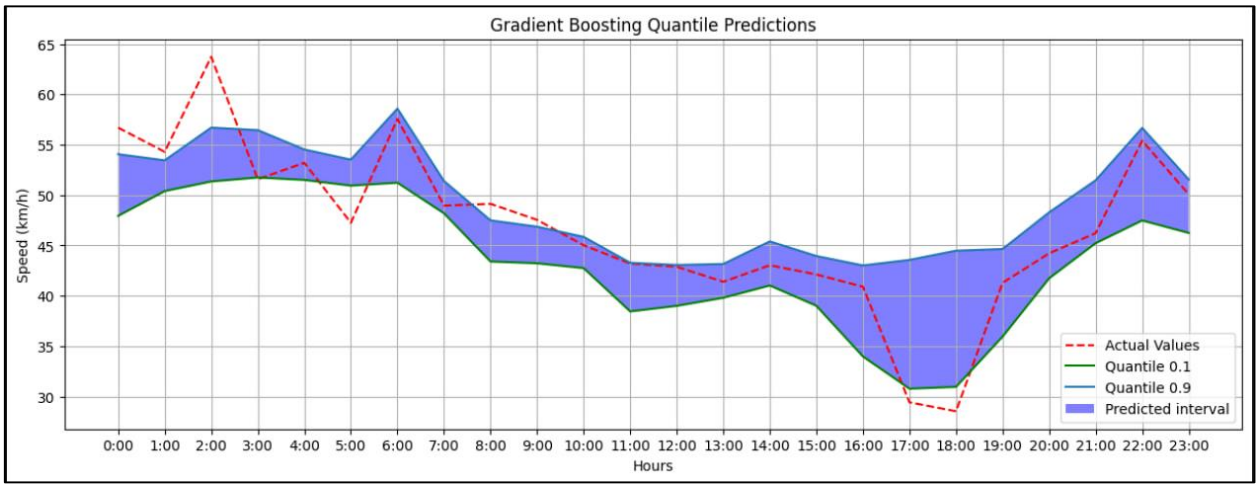

Рисунок 4.7 – Графік квантильної регресії градієнтного бустингу з прогнозованими значеннями на добу (Рисунок виконаний самостійно)

Вміст таблиці 4.8 наведено детальну інформацію про прогнозовані значення за допомогою регресії градієнтного бустингу для всіх квантилів.

Таблиця 4.8 – Фрагмент прогнозованих значення за допомогою методу градієнтного бустингу (Таблиця виконана самостійно)

Фактичне значення	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
56.694	47.933	50.96	54.06
54.296	50.396	52.037	53.226
63.731	51.352	54.881	56.701
51.569	51.743	54.338	56.45
53.206	51.503	52.956	54.528
47.248	50.939	51.736	53.517
57.56	51.228	54.545	58.588
48.95	48.191	48.302	51.395

Вміст таблиці 4.9 показує метрики помилки для прогнозованих значень за допомогою методу градієнтного бустингу для одного сегменту дороги для всіх квантилів.

Таблиця 4.9 – Метрики помилки для прогнозованих значень за допомогою методу градієнтного бустингу (Таблиця виконана самостійно)

Метрика помилки	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
R^2	0.6176	0.7678	0.5642
MAE	4.0267	2.6925	3.4773
MAPE	0.0835	0.0639	0.089
RMSE	24.2395	14.7228	27.6279

Вміст таблиці 4.10 показує усереднені метрики помилки для прогнозованих значень за допомогою методу градієнтного бустингу для всіх сегментів.

Таблиця 4.10 – Метрики помилки для прогнозованих значень за допомогою методу градієнтного бустингу (Таблиця виконана самостійно)

Метрика помилки	Квантиль 0.1	Квантиль 0.5	Квантиль 0.9
R^2	0.3098	0.7747	0.1063
MAE	4.4955	3.0679	6.4237
MAPE	0.0959	0.0887	0.2273
RMSE	44.4816	31.1053	125.3051

Час прогнозування для цієї моделі склав 4.707 с., а загалом для всіх 62 моделей зайняло 4 хв. 31с. Тобто в середньому прогнозування однієї моделі зайняло 4.18 с.

4.5 Аналіз результатів

В результаті експерименту було отримано результати, всі моделі показали схожі результати. А ось за часом помітна вагома різниця найшвидшим виявився метод KNN, а найдовшим градієнтний бустинг.

Метод k-найближчих сусідів зазвичай працює швидше за випадковий ліс у випадках, коли набір даних не дуже великий. Випадковий ліс може мати сотні дерев, що потребує багато часу та обчислювальних ресурсів. Кожне дерево випадкового лісу виконує оцінку значення цільової змінної на основі деякої підмножини ознак, що може призвести до перенавчання, якщо кількість ознак дуже велика.

А ось, квантильна регресія градієнтного бустингу включає в себе обчислення градієнта функції втрати для кожного дерева регресії, що додається до ансамблю. Це може бути дуже обчислювально витратно, особливо якщо мається велика кількість дерев у моделі.

Проте, на невеликих об'ємах даних випадковий ліс може прогнозувати швидше, оскільки він зазвичай має меншу складність та обчислювальну складність, ніж градієнтний бустинг. Однак, зі збільшенням об'єму даних,

різниця в часі виконання між цими моделями може стати менш помітною, або навіть помінятися місцями. Отже, вибір між методом залежить від конкретної задачі, розміру даних та характеристик ознак.

Якщо дивитися на усереднені метрики всі методи краще працювали для медіани та нижнього квантилю 0.1 ніж для верхнього 0.9, але представлена модель сегменту показала результати навпаки вищі.

Найточніші результати квантильної регресії можуть бути досягнені для медіани в тому випадку, якщо дані розподілені більш-менш рівномірно навколо середнього значення. Якщо розподіл даних має високі хвости або низькі хвости, тоді можливо, що для медіани результати будуть менш точними, ніж для інших квантилів. Це пояснюється тим, що медіана є квантилем 0.5, тобто це значення, яке розділяє дані на дві рівні частини. Це означає, що кожен лист у деревах буде мати близько однакову кількість прикладів, що може покращити точність регресії.

А також різниця в метриках може бути пояснена через деякі особливості даних або обраних параметрів моделі. Один з можливих пояснень полягає в тому, що збільшення квантилю знижує кількість прикладів в кожному листі у деревах випадкового лісу, що може знизити точність моделі. Крім того, вибір параметрів моделі (наприклад, кількість дерев у лісі, глибина дерев і т.д.) може впливати на результати регресії за допомогою для різних квантилей.

Також з результатів було помітне велике значення метрики RMSE. Це може бути пов'язано з низькою якістю даних, невідповідністю моделі до даних або недостатньою кількістю змінних, що використовуються для передбачення.

Тож, за результатами можна сказати, що найбільш точним, хоча й не найшвидшим, виявився алгоритм випадкового лісу.

ВИСНОВКИ

У цій роботі було досліджено доцільність використання різних моделей квантильної регресії для короткострокового прогнозування інтервалів швидкості руху Uber у Києві за даними січня, лютого 2020 року на прикладі 62 сегментів дороги. Були порівняні метрики квантильної регресії k-найближчих сусідів, квантильної регресії випадкового лісу, квантильної регресії з градієнтним підсиленням. Найбільш влучним методом було виявлено метод випадкового лісу.

У результаті виконання роботи було описано аналіз предметної галузі, опис основних факторів, що впливають на умови дорожнього руху, поставлене завдання дослідження та описані його етапи, наведено аналіз існуючих методів, що широко засовуються для прогнозування дорожнього трафіку, проведене планування експериментального дослідження, що включає вибір набору даних, вибір методів, а саме регресійних моделей, та метрик для оцінювання, проведене дослідження і описані його результати.

Загалом, робота дає уявлення про потенціал квантильних регресійних моделей для прогнозування інтервалів швидкості руху і може стати основою для майбутніх досліджень і моделювання в цій галузі.

У майбутньому ми плануємо провести дослідження та порівняти ефективність інших методів машинного навчання, які можна застосувати для прогнозування трафіку, наприклад, нейронних мереж. Потрібно розглянути можливість використання різних факторів. Наприклад, ми можемо додати дані про погоду, події в місті, дорожні роботи та аварії, що допоможе визначити ключові стресові точки в містах.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Khaled Shaaban, Mazen Elamin, Mohammed Alsoub, Intelligent Transportation Systems in a Developing Country: Benefits and Challenges of Implementation, Transportation Research Procedia, vol. 55, 2021, pp. 1373-1380, doi: 10.1016/j.trpro.2021.07.122.

2. Kiev traffic index. URL: https://www.tomtom.com/en_gb/traffic-index/kiev-traffic (дата звернення: 25.02.2023).

3. Daunoras J., Bagdonas V., Gargasas V. City transport monitoring and routes optimal management system. Transport. 2008. 23(2). p. 144-149.

4. Zewei Zhou, Ziru Yang, Yuanjian Zhang, Yanjun Huang, Hong Chen, Zhuoping Yu, A comprehensive study of speed prediction in transportation system: From vehicle to traffic, iScience, vol. 25, Issue 3, 18 March 2022, doi: 10.1016/j.isci.2022.103909.

5. L. Pun, P. Zhao and X. Liu, "A Multiple Regression Approach for Traffic Flow Estimation," in IEEE Access, vol. 7, pp. 35998-36009, 2019, doi: 10.1109/ACCESS.2019.2904645.

6. G. Dai, C. Ma and X. Xu, "Short-Term Traffic Flow Prediction Method for Urban Road Sections Based on Space–Time Analysis and GRU," in IEEE Access, vol. 7, pp. 143025-143035, 2019, doi: 10.1109/ACCESS.2019.2941280.

7. Y. Hou, P. Edara and C. Sun, "Traffic Flow Forecasting for Urban Work Zones," in IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 4, pp. 1761-1770, Aug. 2015, doi: 10.1109/TITS.2014.2371993.

8. S. Bielievtsov, I. Ruban, K. Smelyakov and D. Sumtsov, “Network technology for transmission of visual information”, Selected Papers of the XVIII International Scientific and Practical Conference “Information Technologies and Security” (ITS 2018), Kyiv, Ukraine, November 27, 2018. In CEUR Workshop Proceedings, Vol-2318, 2018, pp. 160-175. https://ceur-ws.org/Vol-2318/.

9. K. Smelyakov, P. Dmitry, M. Vitalii and C. Anastasiya, "Investigation of network infrastructure control parameters for effective intellectual analysis," 2018 14th International Conference on Advanced Trends in Radioelecrtronics,

Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 2018, pp. 983-986, doi: 10.1109/TCSET.2018.8336359.

10. O. Lemeshko, M. Yevdokymenko, O. Yeremenko, A. M. Hailan, P. Segeč and J. Papán, "Design of the Fast ReRoute QoS Protection Scheme for Bandwidth and Probability of Packet Loss in Software-Defined WAN," 2019 IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), Polyana, Ukraine, 2019, pp. 1-5, doi: 10.1109/CADSM.2019.8779321.

11. Bing Liu, Tao Zhang, Weicheng Hu, "Intelligent Traffic Flow Prediction and Analysis Based on Internet of Things and Big Data", Computational Intelligence and Neuroscience, vol. 2022, Article ID 6420799, 12 pages, 2022. https://doi.org/10.1155/2022/6420799.

12. Nicholas G. Polson, Vadim O. Sokolov. Deep learning for short-term traffic flow prediction. Transportation Research Part C: Emerging Technologies, Vol. 79, pp. 1-17, 2017, doi: 10.1016/j.trc.2017.02.024.

13. Brian L. Smith, Billy M Williams, R. K. Keith Oswald. Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C: Emerging Technologies, Vol. 10, №4, pp/ 303-321, 2002, doi: 10.1016/S0968-090X(02)00009-8.

14. Koenker, R. (2005). Quantile Regression (Econometric Society Monographs). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511754098.

15. Meinshausen, N. Quantile regression forests. Journal of Machine Learning Research, vol. 7, pp. 983-999, 2006.

16. Kirill Smelyakov, Olha Klochko, Zoia Dudar, “Building Quantile Regression Models for Predicting Traffic Flow”, COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine. In CEUR Workshop Proceedings, Vol-3387, 2023, pp. 117-132. https://ceur-ws.org/Vol-3387/.

17. X. Yang, Y. Yuan and Z. Liu, "Short-Term Traffic Speed Prediction of Urban Road With Multi-Source Data," in IEEE Access, vol. 8, pp. 87541-87551, 2020, doi: 10.1109/ACCESS.2020.2992507.

18. Y. Yu, X. Han, M. Yang and J. Yang, "Probabilistic Prediction of Regional Wind Power Based on Spatiotemporal Quantile Regression," in IEEE Transactions on Industry Applications, vol. 56, no. 6, pp. 6117-6127, Nov.-Dec. 2020, doi: 10.1109/TIA.2020.2992945.

19. Q. Meng, M. Mourshed and S. Wei, "Going Beyond the Mean: Distributional Degree-Day Base Temperatures for Building Energy Analytics Using Change Point Quantile Regression," in IEEE Access, vol. 6, pp. 39532-39540, 2018, doi: 10.1109/ACCESS.2018.2852478.

20. K. Smelyakov, A. Chupryna, D. Sandrkin and M. Kolisnyk, "Search by Image Engine for Big Data Warehouse," 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 2020, pp. 1-4, doi: 10.1109/eStream50540.2020.9108782.

21. H. Ruan, B. Wu, B. Li, Z. Chen and W. Yun, "Expressway Exit Station Short-Term Traffic Flow Prediction With Split Traffic Flows According Originating Entry Stations," in IEEE Access, vol. 9, pp. 86285-86299, 2021, doi: 10.1109/ACCESS.2021.3087658.

22. Gao, C. Zhou, J. Rong, Y. Wang and S. Liu, "Short-Term Traffic Speed Forecasting Using a Deep Learning Method Based on Multitemporal Traffic Flow Volume," in IEEE Access, vol. 10, pp. 82384-82395, 2022, doi: 10.1109/ACCESS.2022.3195353.

23. Lun Zhang, Qiuchen Liu, Wenchen Yang, Nai Wei, Decun Dong, An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction, Procedia-Social and Behavioral Sciences, vol. 96, 6 November 2013, pp. 653-662 doi: 10.1016/j.sbspro.2013.08.076.

24. P. Dell'Acqua, F. Bellotti, R. Berta and A. De Gloria, "Time-Aware Multivariate Nearest Neighbor Regression Methods for Traffic Flow Prediction," in

IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3393-3402, Dec. 2015, doi: 10.1109/TITS.2015.2453116.

25. Deekshetha, H.R., Shreyas Madhav, A.V., Tyagi, A.K. (2022). Traffic Prediction Using Machine Learning. In: Suma, V., Fernando, X., Du, KL., Wang, H. (eds) Evolutionary Computing and Mobile Sustainable Networks. Lecture Notes on Data Engineering and Communications Technologies, vol 116. Springer, Singapore, doi: 10.1007/978-981-16-9605-3_68.

26. Ken Chen, Shasha Zhao and Dengyin Zhang, Short-term Traffic Flow Prediction based on Data-Driven Knearest neighbour Nonparametric Regression, Journal of Physics: Conference Series, vol. 1213, Issue 5, 2019, doi: 10.1088/1742-6596/1213/5/052070

27. Yuan, H., Li, G. A Survey of Traffic Prediction: from Spatio-Temporal Data to Intelligent Transportation. Data Sci. Eng. 6, 63–85 (2021), doi: 10.1007/s41019-020-00151-z.

28. Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C.: Short-term traffic forecasting: Where we are and where we're going. Transport. Res. Part C: Emerg. Technol. 43, 3–19 (2014), doi: 10.1016/j.trc.2014.01.005.

29. Amini, M. R., Feng, Y., Yang, Z., Kolmanovsky, I., & Sun, J. (2020). Long-Term Vehicle Speed Prediction via Historical Traffic Data Analysis for Improved Energy Efficiency of Connected Electric Vehicles. Transportation Research Record, 2674(11), 17–29, doi: 10.1177/0361198120941508

30. Zahid M, Chen Y, Jamal A, Mamadou CZ. Freeway Short-Term Travel Speed Prediction Based on Data Collection Time-Horizons: A Fast Forest Quantile Regression Approach. Sustainability. 2020; 12(2):646, doi: 10.3390/su12020646.

31. Uber Movement. URL: https://movement.uber.com/?lang=en-US (дата звернення: 29.03.2023).

32. Kegler. URL: https://kepler.gl/demo (дата звернення: 31.03.2023).

ПЕРЕЛІК ДЖЕРЕЛЬ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

8. S. Bielievtsov, I. Ruban, K. Smelyakov and D. Sumtsov, “Network technology for transmission of visual information”, Selected Papers of the XVIII International Scientific and Practical Conference “Information Technologies and Security” (ITS 2018), Kyiv, Ukraine, November 27, 2018. In CEUR Workshop Proceedings, Vol-2318, 2018, pp. 160-175. https://ceur-ws.org/Vol-2318/.

9. K. Smelyakov, P. Dmitry, M. Vitalii and C. Anastasiya, "Investigation of network infrastructure control parameters for effective intellectual analysis," 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 2018, pp. 983-986, doi: 10.1109/TCSET.2018.8336359.

20. K. Smelyakov, A. Chupryna, D. Sandrkin and M. Kolisnyk, "Search by Image Engine for Big Data Warehouse," 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 2020, pp. 1-4, doi: 10.1109/eStream50540.2020.9108782.