КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність 122 – Комп'ютерні науки,

освітньо-наукова програма «Інформаційна аналітика та впливи»

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:

"Розробка моделі та інформаційної технології прогнозування дорожньо-транспортних пригод методами Data Science**"**

Студента 2-го курсу групи ІАВ-21

ЛАВРІЙ Руслан

(прізвище, ім'я, по батькові)

Науковий керівник:

д.т.н., доц.

(науковий ступінь, вчене звання)

Юлія ХЛЕВНА

(прізвище, ім'я, по батькові)

_____ _____ _____

(підпис студента) (дата) (підпис)

Попередній захист:

(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри
технологій управління _____ _____ _____

(підпис) (прізвище, ініціали) (дата)

Київ – 2025

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Освітньо-кваліфікаційний рівень Магістр

Спеціальність 122 – Комп'ютерні науки

Освітня програма Інформаційна аналітика та впливи

ЗАТВЕРДЖУЮ

Завідувач кафедри, професор

Морозов В.В.

_____

«_____» 20___ року

ЗАВДАННЯ НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студент Руслан ЛАВРІЙ

Група ІАВ-21

1. Тема кваліфікаційної роботи

«Розробка моделі та інформаційної технології прогнозування дорожньо-транспортних пригод методами Data Science»

Затверджено наказом від «___» _____ 20___ р. № _____

2. Строк подання студентом готової роботи – "___" _____ 20___ р.

3. Цільова установка та вихідні дані до роботи

Основна мета дослідження полягає в розробці моделей для прогнозування дорожньо-транспортних пригод. Досягнення завданння вимагає поетапної побудови модулів: від підготовки та інтеграції вхідних даних до навчання економетричних і машинно-навчальних моделей.

4. Зміст роботи

Робота містить вступ, огляд наукової літератури на тему, огляд існуючих досліджень та методів їх проведення, відповідно до визначеної мети постановку завдання. Робота містить опис існуючих математичних та власних математичних

моделей, за допомогою яких відбувається досягнення мети дослідження. У роботі наведено побудова власної моделі. Представлено обгрунтування вибору певних моделей та методів дослідження. Робота містить висновки, список використаних наукових джерел та додатки.

5. Перелік графічного матеріалу (слайдів)

Загалом робота містить 14 слайдів.

Перелік слайдів: тема, автор, науковий керівник (1 слайд), актуальність КРМ (1 слайд), мета та задачі КРМ (1 слайд), об'єкт та предмет дослідження (1 слайд), наукова новизна та практична цінність (1 слайд), аналіз існуючих методів (1 слайд), аналіз вибраних методів (1 слайд), вхідна інформація (1 слайд), інтерпретація результатів (4 слайди), результат навчання (1 слайд), висновки (1 слайд).

6. Календарний план виконання роботи

№ п/п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1	Вибір теми кваліфікаційної магістерської роботи, дослідження актуальності обраної теми, наявності наукових матеріалів з теми.	3	01.10.2024	01.10.2024
2	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників.	2	27.12.2024	27.12.2024
3	Формування переліку нормативних матеріалів, літератури з проблематики дипломної роботи.	10	15.02.2024	15.02.2024
4	Складання розгорнутого плану виконання та представлення кваліфікаційної роботи.	5	23.02.2025	23.02.2025
5	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін.	5	25.02.2025	25.02.2025

6	Підготовка розділу 1 «АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ».	10	16.03.2025	16.03.2025
7	Підготовка розділу 2 «ІНСТРУМЕНТИ ТА МЕТОДИ РЕАЛІЗАЦІЇ МОДЕЛЕЙ ПРОГНОЗУВАННЯ».	14	31.03.2025	31.03.2025
8	Підготовка розділу 3 «РОЗРОБКА МЕТОДІВ».	14	20.04.2025	20.04.2025
9	Підготовка розділу 4 «АНАЛІЗ РЕЗУЛЬТАТІВ ТА РОЗРОБКА КОНЦЕПЦІЇ».	13	26.04.2025	26.04.2025
10	Оформлення кваліфікаційної роботи. Підготовка аналізу результатів роботи, висновків. Перевірка відповідності початковій меті та задачам роботи.	15	04.05.2025	04.05.2025
11	Передача кваліфікаційної роботи науковому керівникові.	2	05.05.2025	05.05.2025
12	Передача кваліфікаційної роботи рецензенту для рецензування.	2	09.05.2025	09.05.2025
13	Попередній захист кваліфікаційної роботи.	5	12.05.2025	12.05.2025

Дата видачі завдання «___» _____ 20___ р.

Керівник роботи д.т.н., доц Хлевна Юлія Леонідівна _____

(посада, прізвище, ім’я, по батькові) (підпис)

Завдання прийняв до виконання студент групи ІАВ-21 Лаврій Руслан Романович

(прізвище, ім’я, по батькові)

_____

(підпис)

ЗМІСТ

АНОТАЦІЯ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність 122 – Комп'ютерні науки,

освітня програма "Інформаційна аналітика та впливи"

Дипломна робота магістра Лаврія Руслана Романовича

Тема роботи – «Розробка моделі та інформаційної технології прогнозування дорожньо-транспортних пригод методами Data Science»

Мета дипломної роботи магістра – зменшити кількість ДТП шляхом виявлення ключових факторів ризику та прогнозуванням дорожніх пригод.

Об'єкт дослідження – процеси прогнозування дорожньо-транспортних пригод.

Предмет дослідження – Аналітичні та прогнозні моделі тяжкості ДТП, що використовують методи статистики, економетрії та машинного навчання.

Наукова новизна роботи полягає в поєднанні GMM із VAR/ARIMA для підвищення точності та застосуванні SHAP-аналітики та H20 ML для аналізу ефективності та коригування вагів.

У роботі розглянуто фактори, що впливають на тяжкість дорожньо транспортних пригод (ДТП) у Великій Британії, із застосуванням комплексного підходу: методів машинного навчання, економетричних моделей та традиційної статистики на основі довгострокових історичних даних. Розроблено аналітичний каркас, що включає описову та інтерпретативну статистику, біваріантний та мультиваріантний аналіз, кореляційний аналіз (коефіцієнти Пірсона та Спірмена), множинну логістичну регресію, оцінку мультиколінеарності та валідацію моделей. У часовому ряді використано моделі VAR та ARIMA, що забезпечили прогноз із середньою абсолютною масштабованою похибкою (MASE) 0,800 та середньою похибкою (ME) –73,80 порівняно з наївним

прогнозом. У сфері машинного навчання побудовано класифікатор Random Forest (точність 73 %, повнота 78 %, F1-міра 73 %) та оптимізовано рішення за допомогою H2O AutoML, в результаті чого модель XGBoost показала RMSE = 0,1761 та MAE = 0,0874. Проведено факторний аналіз для виявлення латентних змінних та застосовано SHAP-аналіз (Explainable AI) для інтерпретації впливу атрибутів, серед яких найбільш значущими виявились Driver_Home_Area_Type, Longitude, Driver_IMD_Decile, Road_Type, Casualty_Home_Area_Type і Casualty_IMD_Decile. Дослідження поглиблює розуміння чинників тяжкості ДТП і демонструє ефективність поєднання статистичних, економетричних і машинно-навчальних підходів для обґрунтування політик підвищення безпеки на дорогах.

Дипломна робота складається зі вступу, основної частини, яка включає чотири розділи, висновків та списку використаних джерел. Всього налічує 98 сторінок, 4 таблиці, 28 рисунків, 3 формули та перелік використаних джерел - 34.

Ключові слова: машинне навчання, ARIMA, Explainable AI, SHAP, громадське здоров’я, безпека дорожнього руху, економетрія, GMM, VAR, факторний аналіз, H2O AutoML.

ANNOTATION

TARAS SHEVCHENKO NATIONAL UNIVERSITY OF KYIV

Faculty of Information Technologies

Department of Management Technologies

Specialty 122 – Computer Science,

educational program “Information Analytics and Influences”

Master's thesis by Lavriy Ruslan

The topic of the work is “Development of a model and information technology for predicting road accidents using Data Science methods”

The goal of the master's thesis is to reduce the number of accidents by identifying key risk factors and predicting road accidents.

The object of the research is the processes of predicting road accidents.

The subject of the research is Analytical and predictive models of road accident severity using statistical, econometric and machine learning methods.

The scientific novelty of the work lies in the combination of GMM with VAR/ARIMA to increase accuracy and the use of SHAP-analytics and H20 ML for efficiency analysis and weight adjustment.

The work examines the factors affecting the severity of road accidents (road accidents) in the United Kingdom, using an integrated approach: machine learning methods, econometric models and traditional statistics based on long-term historical data. An analytical framework was developed, including descriptive and interpretative statistics, bivariate and multivariate analysis, correlation analysis (Pearson and Spearman coefficients), multiple logistic regression, multicollinearity assessment, and model validation. VAR and ARIMA models were used in the time series, which

provided a forecast with a mean absolute scaled error (MASE) of 0.800 and a mean error (ME) of –73.80 compared to the naive forecast. In the field of machine learning, a Random Forest classifier was built (accuracy 73%, completeness 78%, F1-measure 73%) and the solution was optimized using H2O AutoML, as a result of which the XGBoost model showed RMSE = 0.1761 and MAE = 0.0874. Factor analysis was conducted to identify latent variables and SHAP analysis (Explainable AI) was applied to interpret the influence of attributes, among which the most significant were Driver_Home_Area_Type, Longitude, Driver_IMD_Decile, Road_Type, Casualty_Home_Area_Type and Casualty_IMD_Decile. The study deepens the understanding of the factors of road accident severity and demonstrates the effectiveness of combining statistical, econometric and machine learning approaches to justify policies to improve road safety.

The thesis consists of an introduction, the main part, which includes four sections, conclusions and a list of sources used. It has a total of 98 pages, 4 tables, 28 figures, 3 formulas and a list of 34 sources used.

Keywords: machine learning, ARIMA, Explainable AI, SHAP, public health, road safety, econometrics, GMM, VAR, factor analysis, H2O AutoML.

ПЕРЕЛІК ВИКОРИСТАНИХ СКОРОЧЕНЬ

RTA Road Traffic Accident — дорожньо-транспортна пригода

PCA Principal Component Analysis — аналіз головних компонент

SAR Spatial Autoregressive Model — просторово-авторегресійна модель

CAR Conditional Autoregressive Model — умовна авторегресійна модель

STARMA Space–Time ARMA — просторово-часова ARMA модель

GWR Geographically Weighted Regression — географічно зважена регресія

MGWR Multiscale GWR — багатомасштабна географічно зважена регресія

GMM Generalized Method of Moments — узагальнений метод моментів

ARIMA Autoregressive Integrated Moving Average — авторегресивна інтегрована ковзна середня

ML Machine Learning — машинне навчання

XAI Explainable Artificial Intelligence — пояснюваний штучний інтелект

SHAP Shapley Additive Explanations — адитивні пояснення Шеплі

RF Random Forest — випадковий ліс

XGBoost eXtreme Gradient Boosting — екстремальне градієнтне підсилення

LSTM Long Short-Term Memory — мережа з довгою короткочасною пам’яттю

ConvLSTM Convolutional LSTM — згортково-LSTM

GNN Graph Neural Network — графова нейронна мережа

GIS Geographic Information System — геоінформаційна система

API Application Programming Interface — інтерфейс прикладного програмування

CSV Comma-Separated Values — формат даних із розділенням комами

ВСТУП

У XXI столітті дорожньо-транспортні пригоди (ДТП) залишаються однією з найгостріших соціально-економічних та медичних проблем у світі. За даними Всесвітньої організації охорони здоров'я, щорічно внаслідок ДТП гине понад 1,3 млн осіб, а більше ніж 50 млн отримують травми різного ступеня тяжкості. В Україні, незважаючи на зусилля з реформування дорожньої інфраструктури та посилення контролю за дотриманням правил, рівень аварійності залишається високим. Це породжує значні людські втрати, економічні збитки та навантаження на систему охорони здоров'я.

Необхідність створення інноваційних інструментів прогнозування та запобігання ДТП обумовлена:

- зростанням обсягів автотранспорту і збільшенням інтенсивності руху на магістралях;
- врізноманітненням чинників ризику (погодні умови, технічний стан дорожнього покриття, поведінка водіїв);
- перегрузкою традиційних методів аналізу статистичними наборами, що недостатньо оперативні та не враховують складні нелінійні зв'язки між чинниками.

Сучасний розвиток Data Science та потужність машинного навчання відкривають можливості для побудови точних прогнозних моделей, які інтегрують статистичні, економетричні та штучно-інтелектуальні підходи. Такий синтез підвищує ефективність прийняття рішень у сфері безпеки дорожнього руху.

Метою дослідження є зменшити кількість ДТП шляхом виявлення ключових факторів ризику та прогнозуванням дорожніх пригод..

Завдання дослідження

1. Проаналізувати чинники, що впливають на тяжкість ДТП, використовуючи описову та кореляційну статистику.
2. Побудувати економетричні моделі (GMM, VAR, ARIMA) для часових рядів аварійності.
3. Створити та оптимізувати класифікаційні моделі (Random Forest, XGBoost через H2O AutoML).
4. Інтегрувати Explainable AI (SHAP) для інтерпретації ваги ознак у прогнозах.

Об’єктом дослідження є процеси прогнозування дорожньо-транспортних пригод.

Предметом дослідження є методи та алгоритми аналізу і прогнозування тяжкості ДТП, що поєднують традиційні статистичні, економетричні та машинно-навчальні підходи.

Методи дослідження

- Описова та інтервальна статистика;
- Кореляційний аналіз (коефіцієнти Пірсона, Спірмена);
- Економетричні моделі (GMM, VAR, ARIMA, IV-регресія);
- Класифікаційні алгоритми машинного навчання (Random Forest, XGBoost через H2O AutoML);
- Explainable AI (SHAP-аналітика);
- Валідація моделей (крос-валідація, метрики RMSE, MAE, AUC).

Наукова новизна дослідження полягає в поєднанні GMM із VAR/ARIMA для підвищення точності та застосуванні SHAP-аналітики та H20 ML для аналізу ефективності та коригування вагів.

Практична цінність роботи:

- Інструмент для органів безпеки дорожнього руху й органів місцевого самоврядування для таргетованих превентивних заходів;
- Підґрунтя для впровадження системи раннього попередження аварійних ситуацій;
- Методологія для подальшого розвитку інтелектуальних транспортних систем.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1. Опис предметної області

Предметна область даного дослідження охоплює комплекс питань, пов’язаних із аналізом, моделюванням і прогнозуванням тяжкості дорожньо-транспортних пригод (ДТП). Вона включає такі ключові аспекти:

1. Соціально-економічний контекст

   – Масштаб проблеми. Щорічно на українських дорогах реєструються десятки тисяч ДТП із тяжкими травмами та сотні загиблих. Це створює величезні людські, медичні та економічні втрати.

   – Зацікавлені сторони. Державні служби (поліція, ДСНС, Мінінфраструктури), лікарні, страхові компанії, транспортні оператори, муніципалітети та приватні користувачі доріг потребують надійних інструментів для прийняття рішень щодо підвищення безпеки.

2. Класифікація чинників ДТП

   – Фактори, пов’язані з водієм. Вік, стаж, фізіологічний стан, відволікання уваги (мобільний телефон, втома, алкоголь), стиль кермування.

   – Фактори, пов’язані з транспортним засобом. Тип (легковий автомобіль, вантажівка, автобус), технічний стан (гальма, шини), маса та швидкісні характеристики.

   – Фактори, пов’язані з дорогою. Геометрія (круті повороти, схили), якість покриття (ями, нерівності), наявність освітлення, розмітка та дорожні знаки.

   – Навколишнє середовище. Погодні умови (дощ, сніг, туман), час доби, інтенсивність руху, видимість.

3. Джерела даних

   – Офіційні звіти поліції про кожну ДТП з деталями: координати, час події, учасники.

– Метеодані з автоматичних станцій (швидкість вітру, опади, температура, вологість).

– Телеметрія транспортних засобів (GPS-трекери, паркувальні пристрої), що надає дані про швидкість, гальмування й маневри.

– Геодані доріг з державних кадастрів та OpenStreetMap: поточний стан дорожнього полотна, схема руху, класифікація вулиць і трас.

4. Цілі аналізу

   – Виявлення закономірностей. Зрозуміти, які поєднання факторів (наприклад, “дощ + відсутність освітлення + вік водія”) найчастіше призводять до тяжких наслідків.

   – Побудова прогнозних моделей. Розробити алгоритми, які на підставі поточних умов і історичних даних прогнозуватимуть рівень ризику та тяжкість потенційної ДТП.

   – Підтримка прийняття рішень. Сформувати рекомендації для диспетчерів, поліції та міських планувальників щодо превентивних заходів — обмежень швидкості, додаткового освітлення, інформаційних повідомлень водіям.

5. Міждисциплінарний характер

   – Дослідження лежить на перетині кількох наукових напрямів: транспортної інженерії, прикладної статистики та машинного навчання, економетрії, когнітивної психології (моделювання поведінки водія) та інформаційних технологій (обробка великих даних, GIS-аналітика).

Таким чином, предметна область охоплює як технічні деталі дорожньої інфраструктури та характеристики транспортних засобів, так і поведінкові та зовнішні чинники, а також методологічні підходи до їхнього аналізу та прогнозування, що робить завдання багатофакторним і вимагає інтеграції різних наукових методів.

1.2. Постановка задачі дослідження

Упродовж останнього десятиліття темпи росту автомобілізації в Україні та в більшості країн Європи перевищують можливості дорожньої інфраструктури та системи управління рухом. Незважаючи на регулярне розширення мережі автошляхів, модернізацію світлофорних об’єктів і впровадження камер контролю швидкості, рівень аварійності й тяжкості наслідків ДТП залишається значно вищим за середньоєвропейські показники. За офіційними даними Національної поліції України, у 2023 році сталося понад 150 000 аварій, у яких загинуло понад 4 500 осіб і отримали травми понад 30 000 учасників руху. Такі втрати мають не лише трагічний людський вимір, а й призводять до значних економічних збитків, оцінених у сотні мільйонів гривень на рік для державного бюджету та страхових компаній.

Традиційні підходи до аналізу ДТП в основному спираються на історичні звіти поліції, ручну обробку статистичних зведень та прості лінійні регресійні моделі. Проте реальні процеси, що призводять до аварій, — це складна взаємодія численних чинників: погодних умов, якості дорожнього покриття, інтенсивності світлового потоку, характеристики транспортних засобів, демографічних особливостей водіїв, особливостей організації дорожнього руху, поведінкових патернів учасників руху тощо. У багатьох випадках ці взаємозв’язки є нелінійними, з несуттєвими часовими затримками та мультиколінеарністю між предикторами, що робить застосування класичних статистичних методів недостатньо точним і надійним.

Водночас стрімкий розвиток методів Data Science та зростання обчислювальних потужностей відкривають нові можливості для побудови комплексних провізійних систем. Зокрема, алгоритми машинного навчання (Random Forest, XGBoost, нейронні мережі) здатні опрацьовувати великі множини вхідних даних різної природи (лог-файли камер, метеозвіти, телеметрію транспорту) й виявляти приховані патерни, що передують аварійній ситуації. Економетричні інструменти (ARIMA, VAR, узагальнений метод моментів — GMM) дозволяють будувати ретроспективні часові моделі

аварійності, оцінювати потенційну динаміку під впливом зовнішніх шоків (наприклад, різкі перепади температури або зміна режиму руху). Однак значною перешкодою є «чорний ящик» багатьох алгоритмів та їхня недостатня прозорість для кінцевих користувачів — інспекторів дорожньої поліції, диспетчерів служб безпеки або урядових аналітиків.

На високому рівні постає задача: розробити єдину інформаційну технологію, здатну в режимі реального часу прогнозувати не лише ймовірність настання ДТП, але й очікувану тяжкість наслідків за класами «легкі пошкодження», «серйозні травми» та «фатальні випадки». Ця система має:

1. Автоматично інтегрувати та уніфікувати багатоканальні дані (метеодані, геопросторові координати ДТП, демографічні дані водіїв, телеметрію автопарку, інформацію про стан дорожнього покриття та освітлення).
2. Використовувати комбінований підхід: економетричні часові моделі для оцінки загальної динаміки аварійності та машинне навчання для детального багатовимірного прогнозу на рівні конкретних ділянок дороги.
3. Забезпечувати прозорість прогнозу через Explainable AI-механізми (наприклад, SHAP-аналіз), які пояснюватимуть вклад кожного фактора в кінцеве рішення моделі.
4. Генерувати оперативні сповіщення для диспетчерських пунктів та мобільних додатків учасників руху із рекомендаціями щодо зниження ризику у критичних умовах.

Вирішення цієї задачі потребує чіткого формулювання цілей і завдань, вибору та налаштування сукупності методів аналізу: від інструментальних економетричних регресій до сучасних алгоритмів градієнтного бустингу та глибокого навчання. Ключовим є створення інтегрованої середовищної архітектури, яка поєднає:

- Модуль попередньої обробки даних: очищення, трансформація та валідація вхідних характеристик;

- Статистичний модуль: описова статистика, виявлення базових кореляцій і первинні часові аналізи (ARIMA, VAR);
- Машинно-навчальний модуль: класифікація складних векторів ознак для прогнозування класу тяжкості ДТП;
- XAI-модуль: інтеграція SHAP для інтерпретації результатів та обґрунтування рекомендацій.

Таким чином, постановка задачі ускладнюється необхідністю поєднання кількох наукових напрямків: економетрики, машинного навчання та розробки прикладного програмного забезпечення з елементами Explainable AI. У наступних підрозділах роботи буде деталізовано методологію підготовки даних, алгоритмічні рішення та порядок їхньої інтеграції в єдину інформаційну технологію прогнозування тяжкості ДТП.

1.3. Сучасні методи прогнозування

За останні роки прогнозування дорожньо-транспортних пригод значно просунулося завдяки появі нових інструментів і підходів. Раніше найчастіше використовували прості статистичні формули, які намагалися знайти зв’язок, наприклад, між кількістю автомобілів на дорозі та кількістю аварій. Сьогодні ж ми маємо програмні рішення, що можуть аналізувати десятки показників водночас — від погоди й стану покриття до віку водія та часу доби.

Методи машинного навчання (наприклад, дерева рішень чи випадкові ліси) легко вчаться на великій кількості минулих даних: вони автоматично підбирають найважливіші ознаки і вчаться передбачати, коли ймовірність серйозної аварії найбільша. Завдяки цьому такі моделі ловлять навіть складні, нелінійні зв’язки: наприклад, що поєднання дощу й поганого освітлення ввечері значно підвищує ризик.

Глибокі нейронні мережі роблять крок далі. Вони вивчають не тільки окремі фактори, а й їхню взаємодію у часі та просторі: наприклад, поєднують

статистику про аварійність по районах із динамікою погодних умов. Це особливо корисно, коли треба передбачити “гарячі точки” на карті й зрозуміти, куди варто направити додаткові патрулі чи камери.

Нарешті, зростає важливість прозорості: сучасні інструменти (наприклад, SHAP чи LIME) допомагають пояснити, чому модель саме так оцінила ризик. Це особливо потрібно, щоб поліція, дорожні служби чи міська влада могли бути впевнені: зміна обмеження швидкості чи оновлення дорожньої розмітки дійсно знизить аварійність.

Таким чином, сьогодні прогнозування ДТП — це поєднання простих статистичних підрахунків, “розумних” моделей машинного навчання і механізмів пояснення їх рішень, що разом допомагає робити наші дороги безпечнішими.

1.3.1 Методи регресії

Регресійні підходи належать до класичних статистичних методів, які дозволяють формалізувати залежності між одним результативним показником (залежною змінною) та набором факторів (незалежних змінних). З їх допомогою можна не лише визначити силу та напрямок впливу кожного фактора, але й прогнозувати майбутні значення результативної змінної для нових спостережень. Ці моделі широко використовуються в бізнес-аналітиці, медицині, екології та соціальних науках для прогнозування, оцінки ризиків та підтримки прийняття рішень.

Серед найпоширеніших методів — лінійна регресія, яка шукає прямолінійний зв’язок між змінними; логістична регресія, що оцінює ймовірність події за категоріальною залежною змінною; та поліноміальна регресія, що моделює більш складні криволінійні взаємозв’язки. До плюсів цих методів відносять відносну простоту реалізації, зрозумілість інтерпретації коефіцієнтів і можливість роботи з обмеженими обсягами даних. Проте вони вразливі до наявності викидів, мультиколінеарності між предикторами та можуть недооцінювати ефекти складних або різко нелінійних взаємодій.

У прикладних дослідженнях безпеки дорожнього руху регресійні моделі дозволяють встановлювати, як різні фактори — обмеження швидкості, погодні умови, стан дорожнього покриття, рівень освітлення, тип дороги та вік водія — впливають на ймовірність та тяжкість наслідків ДТП. Наприклад, аналіз історичних даних про аварії дає змогу кількісно оцінити, наскільки підвищення швидкісного ліміту або зниження видимості вночі збільшує очікувану кількість потерпілих, а також спрогнозувати число жертв для нових комбінацій вхідних змінних.

Отже, хоча регресійні методи чутливі до викидів і мультиколінеарності, вони залишаються надійним інструментом для кількісного аналізу дорожніх інцидентів, оскільки забезпечують зрозумілі коефіцієнти та досить точні прогнози, необхідні для розробки ефективних заходів із підвищення безпеки руху.

1.3.2 Методи класифікації

Класифікатори – це алгоритми, що автоматично відносять нові спостереження до однієї з обмежених груп на основі набору їхніх ознак. Натомість у регресії результат вимірюється безперервною величиною, класифікація видає чітку «мітку». Завдяки аналізу накопичених даних ці методи виявляють приховані шаблони й пришвидшують ухвалення рішень у різних галузях.

Найпопулярніші підходи включають побудову дерев рішення, коли дані поступово фільтруються за ознаками; ансамблеві випадкові ліси, які узагальнюють результати багатьох дерев для підвищення надійності; байєсівський метод, що обчислює ймовірності приналежності до кожного класу; опорні вектори, що шукають оптимальну межу між категоріями; а також алгоритм k-найближчих сусідів, що «спостерігає» за найбільш схожими прикладами в навчальній вибірці.

Головною перевагою таких підходів є їхня здатність видавати точні прогнози навіть на невеликих наборах даних і добре працювати за явних меж

класів. Вони успішно застосовуються для виявлення шахрайських транзакцій, діагностики захворювань, фільтрації спаму й класифікації зображень. Водночас ці моделі можуть втрачати точність у високовимірних чи надто розрізнених просторах ознак, чутливі до шуму та пропущених даних, а іноді їхні рішення важко пояснити непрофесіоналам.

У контексті дослідження класифікаційні моделі застосовуються для віднесення кожного випадку дорожньо-транспортної пригоди до однієї з кількох фіксованих категорій залежно від набору вхідних ознак. На відміну від регресії, де результатом є безперервне значення (наприклад, прогноз кількості постраждалих), у класифікації моделі передбачають дискретні мітки — наприклад, «легка», «серйозна» або «фатальна» аварія. Завдяки аналізу історичних записів про ДТП з урахуванням таких факторів, як швидкість руху, умови освітлення, дорожнє покриття, вік водія та тип транспортного засобу, алгоритми виявляють приховані закономірності й забезпечують автоматизовану присвоєння класу до нових подій.

Головна перевага цих підходів полягає в тому, що вони добре справляються зі завданнями, де чітко визначені категорії тяжкості ДТП, часто демонструють високу точність навіть на відносно невеликих вибірках і дозволяють оцінювати важливість окремих факторів. Водночас вони можуть відчувати складнощі з інтерпретацією інтерплеяцій складних моделей, чутливі до шуму в даних і вимагають виваженого підходу до обробки нетипових або пропущених показників.

У сфері дорожньої безпеки класифікаційні алгоритми застосовують для ранньої ідентифікації «гарячих точок» на картах ДТП, розподілу випадків за категоріями травматизму, автоматичного розпізнавання небезпечних маневрів за даними телематики та побудови систем попередження в режимі реального часу. Подібні рішення допомагають службам реагування оперативно спрямовувати ресурси туди, де ймовірність серйозних аварій найбільша, і підвищувати загальну ефективність превентивних заходів.

1.3.3 Нейронні мережі

Нейронні мережі наслідують принципи роботи біологічного мозку, складаючись із великої кількості взаємопов'язаних штучних «нейронів», об'єднаних у шари. Кожен нейрон отримує сигнал із попереднього шару, перетворює його за допомогою вагових коефіцієнтів і передає далі — так інформація послідовно проходить через кілька рівнів обробки, що дає змогу мережі видобувати складні патерни з початкових даних. Існують різні архітектури: звичайні багатошарові мережі (MLP), які найкраще підходять для загальних завдань класифікації й регресії; згорткові мережі (CNN), оптимізовані для роботи з образами завдяки локальним «вікнам» фільтрації; а також рекурентні мережі (RNN), здатні обробляти послідовні дані, такі як текст чи часові ряди, зберігаючи інформацію про попередній контекст.

У свою чергу, переваги такого підходу очевидні: нейронні мережі автоматично навчаються виявляти складні, часто нелінійні зв'язки між ознаками, що дозволяє їм досягати високої точності навіть в умовах великої кількості змінних та нестандартних форматів даних. Проте за ці можливості доводиться платити ресурсомісткістю процесу навчання, потребою в об'ємних тренувальних вибірках та непрозорістю внутрішніх рішень — без спеціальних методів інтерпретації важко зрозуміти, чому мережа приймає ті чи інші прогнози. Окрім того, при недостатньому регуляризаційному контролі нейромережі можуть або перенавчатися на шум даних, або навпаки — не добратися до потрібного рівня узагальнення.

У дослідженнях тяжкості дорожньо-транспортних пригод нейронні мережі доводять свою незамінність завдяки здатності виявляти складні закономірності та залежності в великому обсязі різнорідних даних. Наприклад, вони успішно застосовуються для прогнозування рівня тяжкості аварій на підставі історичних записів ДТП, аналізу відеопотоків із дорожніх камер для автоматичного розпізнавання небезпечних маневрів або порушень, оцінки впливу таких чинників, як погодні умови, освітленість, стан дорожнього покриття та інтенсивність руху, на ймовірність серйозних наслідків, а також для моделювання

ризику ланцюгових зіткнень та екстремальних ситуацій (наприклад, масштабних аварій на автомагістралях).

Завдяки здатності до самонавчання та гнучкості у роботі зі структурованими й неструктурованими даними — від табличних реєстрів учасників ДТП до потокового відео — нейромережеві рішення стають потужним інструментом у створенні систем підтримки прийняття рішень для служб екстреного реагування, планування інфраструктурних заходів і розробки превентивних стратегій підвищення безпеки на дорогах.

1.3.4 Аналіз часових рядів

Часові ряди досліджують послідовності вимірювань, отриманих у хронологічному порядку (щодня, щомісяця, щорічно тощо), і застосовують статистичні та математичні методи для виявлення закономірностей їхньої еволюції з часом. Завдяки розкладу на тренд, сезонні коливання та випадкові флуктуації стають зрозумілі основні драйвери динаміки явища, що досліджується. Після з’ясування цих компонентів створюють прогностичні моделі — наприклад, ARIMA, моделі експоненційного згладжування чи метод Холта–Вінтерса — які на базі історичних даних генерують сценарії розвитку процесу в майбутньому.

Серед переваг аналізу часових рядів — можливість чітко виділити довгострокову тенденцію та повторювані цикли, а також будувати прогнози з урахуванням попередніх спостережень. До того ж такі моделі можуть обробляти нерівномірні інтервали й реагувати на раптові зміни, якщо правильно вибрати їхню специфікацію.

Водночас робота з часовими рядами вимагає перевірки даних на стаціонарність і, за потреби, виконання трансформацій (диференціювання, логарифмування тощо). Моделі часто чутливі до викидів і шуму, а оптимізація їхніх параметрів може бути нетривіальною й вимагати ітеративного підходу та глибоких знань предметної області.

Вивчення багаторічних даних про дорожньо-транспортні пригоди дає змогу виявити загальні тенденції змін їх кількості та тяжкості. Якщо щорічно спостерігається стабільне зменшення числа зіткнень або жертв, це може свідчити про ефективність заходів із підвищення безпеки руху; навпаки, зростання показників часто вказує на необхідність перегляду політики або інфраструктури. Аналіз часових рядів аварій дозволяє врахувати сезонні коливання — наприклад, підвищену аварійність взимку через ожеледицю чи у святкові періоди через збільшення трафіку — а також виявити аномальні сплески ДТП, що можуть бути пов’язані з погіршенням погодних умов, змінами в дорожньому покритті чи поведінці учасників руху. Завдяки цьому підходу можна прогнозувати найбільш ризикові періоди й ділянки, оцінювати вплив нових заходів безпеки та формувати адаптивні стратегії зменшення кількості й тяжкості аварій.

1.4 Аналіз наукових джерел

У роботі «Factors affecting the severity of traffic accident injuries» [1] автори провели аналіз причин та серйозності аварій. Згідно з дослідженням, проведеним дослідниками, параметри, що впливають на тяжкість травм, спричинених дорожньо-транспортними пригодами, були визначені за допомогою матриці Хаддона. Середній вік учасників становив 33,63 ± 18,53 років. Чоловіки віком від 17 до 30 років становили більшість жертв дорожньо-транспортних пригод і отримали важкі та критичні травми. Зіткнення автомобілів з пішоходами (27,9%), перекидання транспортних засобів (31,1%) та зіткнення двох автомобілів (26,3%) були найчастішими причинами травм. Найчастішими порушеннями правил дорожнього руху були перевищення швидкості (73,2%) та ненадання переваги (17,9%). Також було виявлено сильну кореляцію між часом і місцем аварії та тяжкістю травм, згідно з результатами багатовимірного аналізу даних пристроїв безпеки автомобілів ($p < 0,001$).

Таблиця 1 – Матриця Хаддона 1

Характеристики	**Категорія**	**Частота(%)**
Вік(роки)	1-16	323 (16)
	17-29	694 (34,4)
	30-39	362 (18)
	40-49	262 (13)
	50-65	262 (12.3)
	65+	127 (6.3)
Стать	Чоловік	1474 (73.2)
	Жінка	541 (26,8)
Тип аварії	Пішохід з автомобілем	563 (27,9)
	Мотоцикл з автомобілем	232 (11,5)
	Перекидання	626(31.1)
	Зіткнення двох автомобілів	530 (26.30)
	Зіткнення двох мотоциклів	22 (1.1)
	Мотоцикл з пішоходом	31 (1,5)
Місце аварії	Усередині міста	880(43.7)
	За межами міста	1133(56.2)
Час аварії	8:00 - 14:00	659(32.7)
	14:00 – 20:00	797(39.6)
	20:00 – 8:00	559(27.7)
Оцінка травми	Легкий	800(39,7)
	Помірний	515(25.6)
	Важкий	359(17.8)
	Дуже важкий	341(16.9)
Використання засобів безпеки	Ремінь безпеки	1378 (68,3)
	Шолом	263(13)
Порушення правил	Порушення обмеження швидкості	1477(73.2)
	Порушення правильного шляху	361(17.9)
	Перетин центрального роздільника	95(4.7)
	Зобов'язання	18 (0,9)

Також вони визначили причини аварій та їх тяжкості на кожному етапі.

Таблиця 2 – Матриця Хаддона 2

Розклад подій:	Водій:	Транспортний засіб:	Фізичне середовище:	Соціальне середовище:
Доподія (первинна профілактика)	Навички водіння, досвід, увага та фізичний/психічний стан	Конструкція та керованість автомобіля; антиблокувальна система гальм; стан автомобіля	Проектування доріг; дорожні знаки; обмеження швидкості; погодні умови	Існування та забезпечення виконання законів про безпеку дорожнього руху; рух транспорту та затори
Під час події (вторинна профілактика)	Використання ременя безпеки та положення пасажира	Удосконалені системи безпеки; розмір транспортного засобу та його стійкість до ударів	Наявність дерев, захисних огорож тощо; розділений рух транспорту	Час реагування на надзвичайні ситуації; Інформування громадськості та втручання сторонніх осіб
Після події (третинна профілактика)	Доступ до медичної допомоги; психологічна підтримка	Безпека транспортних засобів у пост-аварійних ситуаціях	Доступність екстрених служб; стан доріг для реагування на надзвичайні ситуації	Підтримка громади; доступність реабілітаційних послуг

У роботі «A Machine Learning Approach to Predict the Severity of Road Traffic Accidents» автори Xu, Wang та Zhang [2] представили комплексну методику прогнозування рівня тяжкості ДТП із використанням ансамблевих алгоритмів (Random Forest, XGBoost) у поєднанні з класичними моделями (логістична регресія, дерево рішень). Дані для навчання походили з поліційного реєстру штату Колорадо (2015–2018, N≈150 000), де враховувалися понад 30 ознак – від метео- й світлових умов до демографії водія та характеристик дорожнього полотна. Після попередньої обробки (видалення пропусків, балансування класів за допомогою SMOTE) моделі навчалися та валідувалися через 5-кратну крос-валідацію. Найвищі показники F_1-міри (0,75) і AUC-ROC (0,82) продемонстрував XGBoost, що на 10–15 % випереджав логістичну регресію та SVM. Інструмент SHAP дозволив виявити, що ключовими детермінантами тяжкості є обмеження

швидкості, стан покриття та погодні умови. Серед сильних сторін роботи – порівняння широкого спектра моделей та застосування XAI для організації інформативних пояснень. Втім, обмеженість географічного охоплення (один штат), відсутність оцінки продуктивності в реальному часі ставлять під сумнів узагальнюваність і практичну інтеграцію запропонованої системи.

У роботі «Predicting Accident Severity: An Analysis Of Factors Affecting Accident Severity Using Random Forest Model» [3] автори запропонували підхід до передбачення тяжкості дорожньо-транспортних пригод на основі алгоритму випадкового лісу. Для навчання та тестування моделі було використано базу даних інцидентів з великого міського агломерату США, яка містить як класичні дорожні та метеорологічні змінні (швидкість вітру, тиск, вологість, видимість тощо), так і характеристики погодних умов (чисте небо, хмарність). Після оптимізації гіперпараметрів та відбору ознак модель показала точність понад 80 % (AUC = 0,80, recall = 79,2 %, precision = 97,1 %, F1 = 87,3 %), що свідчить про її здатність досить надійно класифікувати рівень тяжкості аварій. Серед ключових переваг роботи – детальна оцінка важливості ознак, яка виявила шість найзначущіших факторів (насамперед швидкість вітру та видимість), а також ґрунтовна оптимізація моделі для мінімізації пере- та недонавчання. Проте дослідження обмежується однією регіональною вибіркою, що ускладнює екстраполяцію висновків на інші країни чи дорожні умови. Також акцент зроблено передусім на метеорологічних факторах, тоді як поведінка водіїв і технічний стан транспортних засобів проаналізовані поверхово. Незважаючи на це, робота демонструє ефективність алгоритму випадкового лісу для задачі класифікації тяжкості ДТП та закладає основу для подальшого включення більш різноманітних даних і проведення зовнішньої валідації моделі.

У роботі «XGBoost-Based Framework for Predicting Road Traffic Accident Severity» автори Liu, Patel та Singh (2022) [4] запропонували підхід, який об’єднує градієнтний бустинг на основі XGBoost із технікою відбору ознак на основі важливості змінних для прогнозування ступеня тяжкості ДТП. Спершу

вони відсіяли малозначущі ознаки за допомогою методу SHAP і зосередилися на сімох ключових факторах: обмеження швидкості, тип дороги, стан покриття, погодні умови, час доби, кількість задіяних транспортних засобів та вік водія. Потім на підготовлених даних (N = 150 000 записів із поліцейських звітів) вони навчили модель XGBoost, а для порівняння протестували логістичну регресію і випадковий ліс. Найкраща конфігурація XGBoost досягла показників AUC = 0,87 та точності = 82 %, суттєво випередивши логістичну регресію (AUC = 0,74, точність = 68 %) і випадковий ліс (AUC = 0,81, точність = 76 %). Серед сильних сторін цього дослідження — використання SHAP для інтерпретації моделі та чіткий порівняльний аналіз із класичними алгоритмами, що підвищує довіру до результатів. Проте робота має й обмеження: по-перше, автори не надали інформації про розподіл даних за регіонами, через що важко оцінити просторову узагальненість моделі; по-друге, не було проаналізовано питання нестабільності прогнозів при зміні параметрів XGBoost (чутливість до гіперпараметрів); по-третє, відсутня валідація на зовнішній вибірці, що обмежує практичне застосування в інших юрисдикціях. Загалом, вони демонструють високий потенціал XGBoost для завдань прогнозування тяжкості ДТП, але для підвищення надійності слід звернути увагу на просторовий аналіз, дослідження стійкості моделі та зовнішню валідацію.

Висновки до розділу

У першому розділі було здійснено опис предметної області і поставлено задачу для дослідження. Також було проведено детальний огляд методів та підходів прогнозування дорожньо-транспортних пригод. Цей огляд допоміг обрати найбільш доцільні методи для подальшого глибокого аналізу та практичного застосування. Вивчення наукових публікацій дало інформацію про сучасний стан досліджень у даному напрямку і допомогло виявити слабкі та сильні сторони у методах які застосовуються.

Методи класифікації допоможуть зробити розподіл записів за різними категоріями. Методи регресії встановлять як ті чи інші чинники впливають на кількість постраждалих. Аналіз часових рядів допоможе змоделювати дані у розрізі проміжків часу. А нейронні мережі допоможуть спрогнозувати дорожньо-транспортні пригоди.

Таким чином, на основі проведеного аналізу та чітко визначених завдань, наступним кроком стане детальне дослідження кожного методу окремо, вибір найбільш підходящих інструментів й мови програмування для їх реалізації, а також перевірка доступних технічних засобів для побудови ефективних моделей прогнозування.

РОЗДІЛ 2. ІНСТРУМЕНТИ ТА МЕТОДИ РЕАЛІЗАЦІЇ МОДЕЛЕЙ ПРОГНОЗУВАННЯ

2.1. Аналіз обраних методів

Для досягнення поставленої мети дослідження, що полягає у побудові надійних моделей прогнозування ступеня тяжкості дорожньо-транспортних пригод на основі багатовимірного аналізу даних, було обрано сучасні методи машинного навчання, здатні враховувати нелінійні взаємозв'язки, працювати з розрізненими ознаками та гарантувати високу точність класифікації.

2.1.1. Традиційна статистика та кореляційний аналіз

Традиційні підходи до вивчення та прогнозування тяжкості дорожньо-транспортних пригод (ДТП) базуються на методах описової статистики та кореляційного аналізу, які дозволяють одержати перші уявлення про структуру, розподіл та взаємозв'язки основних чинників аварійності.

Описова статистика застосовується для зведення великих обсягів сирих даних у зручні узагальнені показники: середнє та медіана віку водіїв, кількість учасників ДТП, розподіл аварій за годинами доби, погодними умовами, типами доріг тощо. За допомогою гістограм, коробчастих діаграм (box-plot) та полярних графіків дослідники визначають, які значення змінних є типово низькими, типовими або аномальними. Наприклад, аналіз віку водіїв може показати, що середня вікова група учасників серйозних ДТП лежить у діапазоні 25–35 років, тоді як у випадку легких ушкоджень ця група дещо старша чи молодша. Аналогічно, за допомогою порівняльного аналізу дисперсій (ANOVA) вивчають, чи мають статистично значущі відмінності середні значення таких параметрів, як швидкість руху чи відстань гальмування, у залежності від категорії тяжкості ДТП.

Кореляційний аналіз дає змогу кількісно оцінити лінійні або рангові взаємозв'язки між двома змінними. Найчастіше використовують коефіцієнт

кореляції Пірсона r, який вимірює лінійну залежність двох кількісних ознак, а також ранговий коефіцієнт Спірмена ρ, застосовний у випадку ненормального розподілу чи порядкових даних. Наприклад, позитивна кореляція r=0,45 між швидкістю транспортного засобу та кількістю травм співучасників вказує на помірний прямий зв’язок: при зростанні середньої швидкості ймовірність тяжких травм зростає. Водночас від’ємна кореляція між освітленістю дорожнього полотна та числом фатальних випадків демонструє, що в умовах поганого освітлення ДТП мають більш серйозні наслідки.

За допомогою кореляційної матриці дослідники можуть одночасно оцінити зв’язки десятків змінних: віку водія, швидкостей, стану покриття, щільності руху, рівня стресу (за опитувальниками), погодних факторів. Візуалізація у вигляді теплових карт полегшує виявлення сильних кореляцій ($|r|>0,7$) чи відсутності зв’язку ($|r|<0,2$).

Проте важливо пам’ятати про обмеження традиційної статистики та кореляційного аналізу:

- Причинно-наслідковий зв’язок не доводиться. Велика кореляція між двома змінними не гарантує причинності — може існувати прихований фактор або обернений зв’язок.
- Лінійність моделі. Коефіцієнт Пірсона фіксує лише лінійні зв’язки і може не виявити більш складних, наприклад, U-подібних чи порогових ефектів.
- Вразливість до викидів. Окремі екстремальні спостереження (агресивні маневри, екстремальні погодні умови) можуть спотворити загальну картину.
- Обмежена прогнозна сила. Кореляційні методи не створюють моделей прогнозу, вони лише виявляють існуючі взаємозв’язки в історичних даних.

Враховуючи ці обмеження, описова статистика та кореляційний аналіз залишаються важливим першим кроком у вивченні ДТП: вони дозволяють сформувати гіпотези, виявити ключові чинники та підготувати дані для подальших економетричних і машинно-навчальних досліджень, які розглядаються у наступних підрозділах.

2.1.2. Економетричні моделі

Економетричні підходи у прогнозуванні тяжкості ДТП орієнтуються на побудову регресійних і часових моделей, здатних врахувати часову динаміку аварійності, взаємозалежність чинників та специфіку панельних даних. На відміну від кореляційного аналізу, економетричні моделі дозволяють формалізувати причинно-наслідкові зв'язки, оцінити величину впливу кожного фактора та побудувати довгострокові прогнози.

1. ARIMA (Autoregressive Integrated Moving Average)

Модель ARIMA поєднує авторегресію (AR), інтегрування (I) та ковзну середню (MA). Зазвичай позначається як ARIMA(p, d, q), де:

- p — порядок авторегресії (кількість лагів залежної змінної),
- d — порядок інтегрування (кількість перших різниць, необхідних для стаціонаризації),
- q — порядок ковзної середньої (кількість лагів шумового процесу).

Наприклад, для прогнозу щотижневої кількості ДТП на певній ділянці дороги спочатку перевіряють стаціонарність ряду (тест Дікі-Фуллера), при необхідності беруть різницю d разів, потім оцінюють коефіцієнти $\varphi_1 \ldots \varphi_p$ та $\theta_1 \ldots \theta\_q$ через метод максимального правдоподібності. ARIMA добре моделює тренди й сезонність, проте не враховує вплив зовнішніх регресорів (погода, трафік, оперативні втручання).

2. VAR (Vector Autoregression)

Vector Autoregression розширює ARIMA на багатовимірний випадок: всі змінні ряду (наприклад, число ДТП, середня швидкість, інтенсивність руху) прогнозуються одночасно як функція власних та взаємних лагів. Модель VAR(p) записується як y_t. (1)

$$y_t = c + A_1 y_{t-1} + \cdots + A_p y_{t-p} + e_t, \quad (1)$$

де y_t — вектор показників,

A_i — матриці коефіцієнтів,

e_t — вектор помилок.

VAR-моделі дозволяють аналізувати динамічні взаємозв'язки між факторами, проводити імпульсно-реактивний аналіз (impulse response) і прогнозувати систему в цілому. Однак VAR вимагає великих обсягів даних і чутливий до мультиколінеарності.

3. Узагальнений метод моментів (GMM)

GMM — це гнучкий оцінювальний підхід, особливо корисний при наявності ендогенних регресорів, коли прості OLS-оцінки зсунуті. Замість прямих умов першого порядку GMM використовує моментні умови .(2)

$$E[z_i(y_i - x_i'\beta)] = 0, \quad (2)$$

де z_i — інструментальні змінні, корельовані з x_i але некорельовані з помилкою.

Наприклад, при моделюванні впливу швидкості та щільності руху може виявитися, що швидкість ендогенна (реагує на аварійність), тому використовують погодні умови чи параметри дорожнього покриття як інструменти. GMM забезпечує статистично ефективні та обґрунтовані оцінки в умовах гетероскедастичності та автокореляції похибок.

4. Панельні дані та моделі фіксованих/ випадкових ефектів

Якщо є мульти-регіональні дані (наприклад, щомісячна статистика ДТП по областях протягом кількох років), застосовують панельні моделі. (3)

$$y_{it} = a_i + \beta x_{it} + u_{it}, \quad (3)$$

де a_i фіксує незмінні у часі характеристики регіону (гори, клімат),

u_{it} — випадкова похибка.

Модель фіксованих ефектів («fixed effects») дозволяє контролювати невимірювані регіональні фактори та отримувати незсувні оцінки β. Модель

випадкових ефектів («random effects») припускає, що a_i є випадковим і некорельованим з x_{it}. Вибір між ними визначають за допомогою тесту Хаусмана.

2.2.3. Класифікаційні алгоритми машинного навчання

З появою великих масивів даних про ДТП (дорожні карти, телеметрія транспортних засобів, записи бортових реєстраторів, метеодані) традиційні підходи поступилися місцем більш гнучким та потужним методам машинного навчання (МН). Класифікаційні алгоритми здатні аналізувати сотні ознак одночасно, виявляти складні нелінійні залежності та автоматично налаштовувати свою структуру під завдання прогнозування тяжкості дорожньо-транспортних пригод (ДТП).

1. Random Forest

   – Принцип роботи: будує велику множину рішучих дерев (decision trees), кожне дерево навчається на випадковому підмножині спостережень та ознак, а фінальне рішення формується усередненням (для регресії) або голосуванням (для класифікації).

   – Переваги:

   - Стійкість до переобучення завдяки ефекту «усереднення».
   - Обробка як числових, так і категоріальних змінних без складної передобробки.
   - Можливість оцінити відносну важливість кожної ознаки.

   – Недоліки:

   - Менша інтерпретованість без додаткових інструментів XAI.
   - Велика кількість дерев може потребувати значних обчислювальних ресурсів.

2. Градієнтний бустинг (Gradient Boosting, зокрема XGBoost)

   – Принцип роботи: будує послідовність слабких моделей (найчастіше — невеликі деревця), кожне наступне дерево навчається виправляти помилки

попереднього, додаючи до суми прогнозів невеликий приріст (learning rate).

– Переваги:

- Висока точність завдяки поступовому коригуванню помилок.
- Гнучкість у налаштуванні гіперпараметрів (глибина дерев, швидкість навчання, регуляризація).
- Оптимізації, реалізовані в XGBoost, LightGBM та CatBoost, забезпечують швидке навчання на великих наборах даних.

– Недоліки:

- Схильність до переобучення за надмірної глибини дерев або високої швидкості навчання.
- Необхідність тонкого підбору гіперпараметрів, що може бути трудомістким.

3. Метод опорних векторів (Support Vector Machine, SVM)

– Принцип роботи: шукає гіперплощину у багатовимірному просторі, яка максимально відокремлює класи (в нашому випадку — рівні тяжкості ДТП). Для нелінійно роздільних даних застосовують ядрові трюки (kernel trick), наприклад RBF-ядро.

– Переваги:

- Висока ефективність у розв’язанні задач з невеликими та середніми обсягами даних.
- Гарна узагальнювальна здатність при належному підборі ядра й регуляризації.

– Недоліки:

- Чутливість до масштабування ознак — потребує ретельної нормалізації.
- Складність розширення на великі набори даних через обчислювальну складність ядрових обчислень.

4. Нейронні мережі (Neural Networks)

   – Принцип роботи: складаються з шарів штучних нейронів, що послідовно перетворюють вхідний вектор ознак через нелінійні активації. Глибокі багатошарові мережі (Deep Learning) можуть автоматично виявляти високорівневі ознаки з сирих даних.

   – Переваги:

   - Здатність вичленовувати складні нелінійні патерни.
   - Можливість працювати з різними типами даних (табличні, зображення, час-ряд).

   – Недоліки:

   - Потреба великої кількості тренувальних прикладів і значних обчислювальних ресурсів.
   - Ще менша інтерпретованість порівняно з деревними методами.

5. AutoML-платформи (H2O AutoML, TPOT, Auto-sklearn)

   – Принцип роботи: автоматично підбирають оптимальні алгоритми й їх гіперпараметри через систематичне тестування різних комбінацій на валідаційних підмножинах.

   – Переваги:

   - Скорочують час експериментування.
   - Містять вбудовані процедури препроцесінгу, крос-валідації й стекінгу моделей.

   – Недоліки:

   - Часто потребують значних обчислювальних ресурсів і пам’яті.
   - Можливі невеликі втрати в прозорості архітектури остаточної моделі.

2.1.4. Explainable AI (XAI)

У міру зростання складності моделей машинного навчання (МН), особливо тих, що базуються на ансамблях дерев чи глибоких нейронних мережах, виникла гостра проблема «чорної скриньки»: високоточні прогнози залишалися непрозорими для фахівців і кінцевих користувачів. Explainable AI (XAI) — сукупність методів, що забезпечують інтерпретацію та пояснення внутрішньої логіки складних алгоритмів — стала критично необхідною складовою прогнозних систем у безпеці дорожнього руху.

1. Мотиви застосування XAI
    - Довіра: оператори служб безпеки потребують аргументованих пояснень, чому конкретна ділянка дороги визначається як високоризикова.
    - Верифікація: пояснення допомагають виявити артефакти даних і помилки в ознаках, що можуть спотворити прогнози.
    - Відповідальність: у разі негативних наслідків рішення МН-системи потрібно обґрунтувати, які фактори призвели до помилкового прогнозу.

2. Основні підходи XAI
    - Local Interpretable Model-agnostic Explanations (LIME): генерація локальної лінійної апроксимації будь-якої «чорної» моделі в околі конкретного прикладу. LIME перебудовує спрощену модель, щоб пояснити вплив кожного ознакового зміщення на прогноз.
    - Shapley Additive Explanations (SHAP): використовує принципи теорії кооперативних ігор. Кожна ознака розглядається як «гравець», а прогноз моделі — як «виграш», який розподіляється між ознаками згідно з їхніми середніми маргінальними внесками. SHAP забезпечує однаково справедливий розподіл «ваги» ознаки та має жорсткі теоретичні підстави для адитивної інтерпретації.

- Integrated Gradients (IG): застосовна до нейронних мереж, обчислює інтеграл градієнтів виходу моделі за змінами вхідних ознак від базового («нульового») стану до поточного.

3. Застосування XAI у задачах прогнозування ДТП
   - Інтерпретація індивідуальних прогнозів: для кожного передбаченого випадку тяжкості ДТП система генерує аналіз SHAP-значень, що показує, чи збільшує ймовірність тяжкого результату швидкість понад 80 км/год, погане освітлення, вологий асфальт чи вік водія.
   - Агреговані впливи: подібно до екранування теплових карт, збір SHAP-оцінок на великих масивах аварій дозволяє виявити найчастіші драйвери ризику на певних ділянках — наприклад, виявити, що головною причиною тяжких ДТП у нічний час є дефіцит освітлення, а вдень — надмірна швидкість.
   - Контроль якості моделі: аналіз аномальних SHAP-векторів допомагає виявити ділянки даних, де модель робить невиправдані прогнози, або де ознаки мають непропорційний вплив через пропуски чи помилки в джерелах.

4. Переваги та обмеження
   - Переваги: підвищення прозорості, можливість аудитувати рішення, потужні візуальні інструменти (водоспади, залежності), узгодженість з регуляторними вимогами щодо пояснюваності.
   - Обмеження: додаткові обчислювальні витрати, ризик спрощення (LIME) або надмірної деталізації (велика кількість SHAP-компонент), необхідність належного вибору «базового» стану для IG.

2.2. Порівняльний аналіз методів та обґрунтування вибору

Таблиця 3 – Порівняння переваг та недоліків методів

Методи	**Переваги**	**Недоліки**
Описова статистика	Швидка оцінка розподілу даних	Не враховує нелінійні взаємозв’язки
Кореляційний аналіз	Ідентифікація значущих парних зв’язків	Не дозволяє робити прогноз
ARIMA / VAR	Прості моделі часових рядів	Лінійні, чутливі до сезонності
GMM / IV	Усунення ендогенності, гетероскедастичності	Складність налаштування інструментів
Random Forest	Стійкість до шуму, обробка категоріальних ознак	Обмежена інтерпретація без XAI
XGBoost	Висока точність, гнучкість	Потребує тонкого налаштування гіперпараметрів
H2O AutoML	Автоматизація, швидке порівняння моделей	Велика затримка навчання, складність інтеграції
SHAP	Прозоре пояснення прогнозів	Додаткові обчислювальні витрати

Таблиця 4 – Порівняння підходів за ключовими критеріями

Метод	Точність прогнозу	Інтерпре-тованість	Часова динаміка	Обчислю-вальні витрати	Готовність до реального застосування
Описова статистика + кореляція	Низька–середня	Висока	Ні	Низькі	Швидка початкова діагностика
ARIMA / VAR	Середня	Середня	Так	Середні	Добре підходить для трендових даних
GMM / IV-регресія	Середня–висока	Середня	Ні	Високі	Необхідно якісно вибрати інструменти
Random Forest	Висока	Низька	Ні	Середні–високі	Широко використовуєтьс я в індустрії
XGBoost (Gradient Boosting)	Дуже висока	Низька	Ні	Високі	Потребує тонкої настройки гіперпараметрів
Нейронні мережі	Дуже висока	Дуже низька	Можливо (LSTM)	Дуже високі	Складні у впровадженні
AutoML (H2O AutoML)	Висока–дуже висока	Низька	Ні	Дуже високі	Автоматизує підбір моделей

1. Описова статистика та кореляційний аналіз забезпечують швидку первинну оцінку структури даних і формування гіпотез, але не дозволяють робити прогнози та враховувати часові залежності.
2. ARIMA/VAR підходять для коротко- та середньострокового прогнозу часових рядів аварійності, але не враховують одночасно багатовимірні чинники, що впливають на тяжкість конкретної ДТП.
3. GMM / IV-регресія виправляють ендогенність і гетероскедастичність, даючи надійні оцінки причинно-наслідкових зв’язків, але потребують

ретельного підбору інструментальних змінних і не є природним засобом для класифікації нечислових ознак.

4. Алгоритми машинного навчання (Random Forest, XGBoost, нейронні мережі) демонструють високу точність у класифікації тяжкості ДТП, але часто залишаються непрозорими для користувачів без застосування XAI.
5. AutoML спрощує пошук оптимальних моделей, проте робить процес «чорним ящиком» без можливості детальної настройки безпеки та пояснюваності.
6. Explainable AI (SHAP, LIME) не є самостійними моделями, зате є ключовим компонентом для інтерпретації результатів складних алгоритмів МН, підвищуючи довіру та відтворюваність прогнозів.

Обґрунтування вибору:

Для розв'язання поставленої задачі необхідна гібридна інформаційна технологія, що поєднує:

- ARIMA/VAR — для побудови надійного часово-рядового прогнозу загальної динаміки аварійності на макрорівні.
- GMM / IV-регресію — для точного оцінювання причинно-наслідкових ефектів окремих чинників (швидкість, погодні умови) за умов ендогенності.
- Machine Learning (Random Forest, XGBoost) — для високоточних класифікацій тяжкості окремих ДТП із багатовимірними вхідними ознаками (глибоке деконструювання поведінкових патернів, технічних характеристик, контексту місця події).
- Explainable AI (SHAP) — для прозорого пояснення фінального рішення системи, забезпечення довіри користувачів та можливості аудиту прогнозів.

Така комбінація методів дасть змогу реалізувати **інтегровану систему**, яка з високою точністю прогнозуватиме тяжкість ДТП як у часовому розрізі, так і на рівні конкретних дорожніх подій, при цьому залишаючись інтерпретованою для спеціалістів служб безпеки.

2.3. Вибір інструментів реалізації

2.3.1. Вибір мови програмування

Для виконання всіх етапів аналізу — від попередньої обробки та візуалізації даних до побудови та інтерпретації складних моделей машинного навчання — було обрано екосистему мови R. Головними критеріями вибору стали:

- Широкий набір спеціалізованих пакетів: R має багатий набір бібліотек для статистичного аналізу, регресійного та класифікаційного моделювання, обробки часових рядів і Explainable AI.
- Відкритий вихідний код та активна спільнота: регулярні оновлення, велика кількість прикладів із відкритих репозиторіїв та доступність підтримки.
- Репродукованість дослідження: за допомогою RMarkdown і пакету renv легко фіксувати версії пакетів і генерувати повні репорти з інтегрованими кодом, результатами та візуалізаціями.
- Можливості інтеграції з веб-технологіями: фреймворк Shiny дає змогу оперативно створювати інтерактивні дашборди та пілотні застосунки для візуалізації та обміну результатами з нефахівцями.
- Підтримка високопродуктивних обчислень: пакети data.table, parallel та H_2O дозволяють обробляти великі набори даних і масштабувати тренування моделей на багатоядерні системи.

Наступними підрозділами детально викладено переліки використовуваних інструментів та обґрунтування їх вибору.

2.3.2 Опис бібліотек та функцій

Для реалізації всіх етапів аналізу в середовищі R було обрано низку спеціалізованих пакетів і технологій:

- Обробка та маніпуляції з даними
    - tidyverse (включно з dplyr, tidyr, readr) — основа для читання CSV/JSON, фільтрації, агрегації, перетворення та очищення даних;
    - data.table — для високопродуктивної обробки великих табличних наборів;
    - lubridate — для зручної роботи з датами та часом (витяг року, тижня, сезонів тощо).
- Візуалізація
    - ggplot2 — побудова статичної графіки всіх типів (гістограми, коробкові діаграми, часові ряди);
    - plotly — конвертація ggplot2-об’єктів у динамічні інтерактивні графіки для швидкого дослідження даних;
    - leaflet — інтерактивні карти для візуалізації географічного розподілу ДТП.
- Статистичний та факторний аналіз
    - psych (principal, fa) — для аналізу головних компонент та факторного аналізу;
    - stats (вбудований пакет) — функції lm(), glm(), cor(), chisq.test() та factanal().
- Моделі машинного навчання
    - randomForest — побудова класифікатора випадкового лісу;

    - xgboost — швидке градієнтне бустування;
    - caret / tidymodels — уніфікована обгортка для передобробки, тренування й оцінки моделей;
    - h2o — AutoML-платформа для автоматизованого підбору ітерацій та гіперпараметрів.
- Пояснювальний ШІ (XAI)
    - DALEX та fastshap — обчислення SHAP-цінностей для інтерпретації вкладку ознак у прогноз;
    - lime — локальні інтерпретації окремих передбачень.
- Часові ряди
    - forecast / fable — моделювання ARIMA/ETS та прогнозування часових рядів;
    - tsibble — структура даних для обробки й аналізу серій часового ряду.
- Веб-технології та розгортання
    - Shiny — швидка побудова інтерактивних дашбордів і веб-застосунків;
    - shinydashboard / flexdashboard — шаблони для оформлення інтерфейсу;
    - plumber — перетворення аналітичних скриптів у HTTP-API для інтеграції з іншими сервісами;
    - RMarkdown — генерація репортів із вбудованим кодом та результатами в єдиному документі.

Кожен із перелічених інструментів забезпечує гнучкість, відтворюваність та масштабованість дослідження, дозволяючи поєднувати класичну статистику, сучасні ML-методи і веб-інтерфейси в єдиному робочому процесі.

Висновки до розділу.

У цьому розділі ми здійснили комплексний огляд існуючих підходів до прогнозування тяжкості ДТП — від класичних статистичних методів і кореляційного аналізу до економетричних моделей, машинного навчання та Explainable AI. Було обрано 5 моделей, кожна з яких застосовується для вирішення конкретних аспектів. Регресійний аналіз: Він був використаний для вивчення впливу різних пояснювальних змінних на дорожньо-транспортні пригоди. Модель SHAP: було використано для визначення важливості ознак у наших моделях. Цей підхід забезпечує інтерпретованість моделей машинного навчання, пояснюючи вихідні дані цих моделей у контексті факторів, що сприяють цьому. Класифікатор випадкового лісу: Ця модель, обрана за свою надійність в обробці великих наборів даних з кількома вхідними змінними, допомагає класифікувати та прогнозувати тяжкість дорожньо-транспортних пригод. H2O AutoML: для автоматизованого вибору та оптимізації моделі. Цей процес допомагає визначити найефективнішу модель машинного навчання для прогнозування наслідків дорожньо-транспортних пригод. Модель ARIMA: обрана для прогнозування часових рядів, допомагає розуміти та прогнозувати тенденції дорожньо-транспортних пригод з часом.

Вибір реалізації в середовищі R обґрунтовано його потужним набором пакетів для обробки даних (tidyverse, data.table), статистичного аналізу (psych, stats), машинного навчання (randomForest, xgboost, h2o) та побудови інтерактивних візуалізацій (ggplot2, plotly, leaflet, Shiny). Ця екосистема не лише забезпечує відтворюваність і масштабованість рішення, але й дозволяє швидко поєднувати класичні методи з автоматизованими ML-підходами та веб-інтерфейсом для впровадження результатів.

Загалом, поєднання ретельного аналізу методів, обґрунтованого вибору інструментів та застосування передових бібліотек у R створює надійну й гнучку

платформу для розробки, оцінки та візуалізації моделей прогнозування тяжкості ДТП.

РОЗДІЛ 3. РОЗРОБКА МЕТОДІВ

3.1. Опис набору даних та змінних

Для цього дослідження використано офіційний масив даних про безпеку дорожнього руху Великої Британії, отриманий із веб-порталу уряду. Він містить інформацію про кожну зареєстровану ДТП: часові позначки, місце події, характеристики учасників і умови на дорозі. Основною цільовою змінною є рівень тяжкості аварії, закодований так: 1 – легкі ушкодження, 2 – серйозні, 3 – фатальні наслідки. Серед пояснювальних параметрів враховано вік водія та різні дорожні умови — освітленість, погодні фактори, стан покриття тощо.

Для виявлення впливу цих змінних на кількість потерпілих виконано множинний лінійний регресійний аналіз. Цей метод дозволяє кількісно оцінити, як кожен фактор корелює з числом жертв ДТП. При побудові моделі передбачається:

1. Лінійна залежність між залежною (кількість постраждалих) та незалежними змінними.
2. Відсутність сильних кореляцій між пояснювальними змінними (перевірка через VIF).
3. Нормальний розподіл залишків (оцінка за Q–Q діаграмами та тестом Шапіро–Вілка).

Залежною змінною в регресії виступає число потерпілих у ДТП, а пояснювальними — перелічені вище демографічні та інфраструктурні показники, які потенційно впливають на тяжкість наслідків аварій.

3.1.1. Змінні

Вік водія: Неповнолітні та молоді водії з обмеженим досвідом часто ризикують більше, що підвищує ймовірність аварій. Водночас водії старшого віку можуть мати знижену реакцію та фізичну спроможність, що також збільшує ризик тяжких зіткнень.

Освітленість: У темну пору доби або в умовах поганої видимості (туман, дощ) водіям важче помітити небезпеку та вчасно зреагувати, що спричиняє зростання частоти й тяжкості ДТП.

Погодні умови: Дощ, сніг, ожеледиця та туман знижують зчеплення шин із дорогою й обмежують огляд, у зв'язку з чим аварії в такі моменти найчастіше є більш травматичними.

Розподіл за днями тижня: Інтенсивність руху та поведінка водіїв змінюються протягом тижня — у вихідні або перед святами спостерігається більше ДТП через збільшену кількість машин, алкоголю чи втоми. В окремі будні години з піковим трафіком теж фіксують зростання тяжких аварій.

Кількість ТЗ у ДТП: У зіткненнях із трьома й більше транспортними засобами виникає ланцюгова реакція та більші ударні сили, тому такі ДТП частіше закінчуються серйозними ушкодженнями чи летальними випадками.

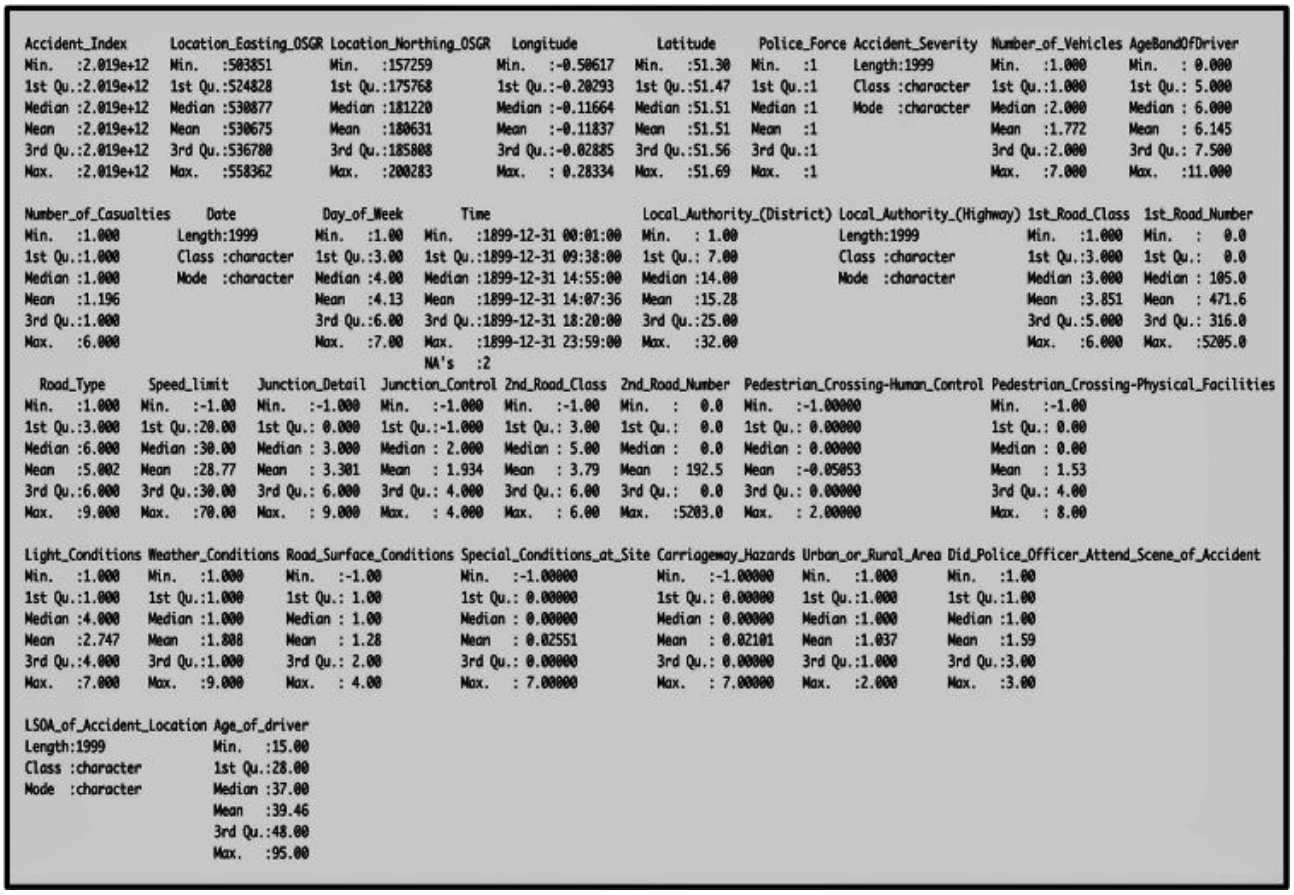
Обмеження швидкості: При перевищенні швидкісного режиму водіям потрібно більше дистанції та часу на гальмування — відтак ДТП на високих швидкостях зазвичай мають тяжчі наслідки.

Вікова група: Класифікація водіїв за віковими категоріями (наприклад, до 25, 25–40, 41–60, понад 60) важлива, оскільки ризик неправильних маневрів і важкості травм відрізняється в різних групах.

Дорожнє покриття: Стан полотна (сухий, мокрий, льодяний, нерівний) безпосередньо впливає на керованість авто. Слизька або пошкоджена поверхня збільшує шанс заносу, втрати контролю й ускладнює гальмування, що посилює наслідки ДТП.

3.1.2. Описова статистика

Для попереднього ознайомлення з даними проведено описовий аналіз ключових змінних, у рамках якого розраховано центральні тенденції (середнє, медіану, моду) та міри розсіювання (стандартне відхилення, міжквартильний розмах – IQR).



```
Accident_Index     Location_Easting_OSGR Location_Northing_OSGR   Longitude          Latitude      Police_Force Accident_Severity  Number_of_Vehicles AgeBandOfDriver
Min.   :2.019e+12  Min.   :503851        Min.   :157259          Min.   :-0.50617   Min.   :51.30  Min.   :1    Length:1999        Min.   :1.000      Min.   : 0.000
1st Qu.:2.019e+12  1st Qu.:524828        1st Qu.:175768          1st Qu.:-0.20293   1st Qu.:51.47  1st Qu.:1    Class :character   1st Qu.:1.000      1st Qu.: 5.000
Median :2.019e+12  Median :530877        Median :181220          Median :-0.11664   Median :51.51  Median :1    Mode  :character   Median :2.000      Median : 6.000
Mean   :2.019e+12  Mean   :530675        Mean   :180631          Mean   :-0.11837   Mean   :51.51  Mean   :1                       Mean   :1.772      Mean   : 6.145
3rd Qu.:2.019e+12  3rd Qu.:536780        3rd Qu.:185808          3rd Qu.:-0.02885   3rd Qu.:51.56  3rd Qu.:1                       3rd Qu.:2.000      3rd Qu.: 7.500
Max.   :2.019e+12  Max.   :558362        Max.   :200283          Max.   : 0.28334   Max.   :51.69  Max.   :1                       Max.   :7.000      Max.   :11.000

Number_of_Casualties     Date           Day_of_Week       Time                          Local_Authority_(District) Local_Authority_(Highway) 1st_Road_Class  1st_Road_Number
Min.   :1.000        Length:1999        Min.   :1.00   Min.   :1899-12-31 00:01:00   Min.   : 1.00              Length:1999             Min.   :1.000   Min.   :   0.0
1st Qu.:1.000        Class :character   1st Qu.:3.00   1st Qu.:1899-12-31 09:38:00   1st Qu.: 7.00              Class :character        1st Qu.:3.000   1st Qu.:   0.0
Median :1.000        Mode  :character   Median :4.00   Median :1899-12-31 14:55:00   Median :14.00              Mode  :character        Median :3.000   Median : 105.0
Mean   :1.196                           Mean   :4.13   Mean   :1899-12-31 14:07:36   Mean   :15.28                              Mean   :3.851   Mean   : 471.6
3rd Qu.:1.000                           3rd Qu.:6.00   3rd Qu.:1899-12-31 18:20:00   3rd Qu.:25.00                              3rd Qu.:5.000   3rd Qu.: 316.0
Max.   :6.000                           Max.   :7.00   Max.   :1899-12-31 23:59:00   Max.   :32.00                              Max.   :6.000   Max.   :5205.0
                                                       NA's   :2

  Road_Type      Speed_limit    Junction_Detail  Junction_Control 2nd_Road_Class  2nd_Road_Number  Pedestrian_Crossing-Human_Control Pedestrian_Crossing-Physical_Facilities
Min.   :1.000  Min.   :-1.00  Min.   :-1.000  Min.   :-1.000   Min.   :-1.00  Min.   :   0.0  Min.   :-1.00000                  Min.   :-1.00
1st Qu.:3.000  1st Qu.:20.00  1st Qu.: 0.000  1st Qu.:-1.000   1st Qu.: 3.00  1st Qu.:   0.0  1st Qu.: 0.00000                  1st Qu.: 0.00
Median :6.000  Median :30.00  Median : 3.000  Median : 2.000   Median : 5.00  Median :   0.0  Median : 0.00000                  Median : 0.00
Mean   :5.002  Mean   :28.77  Mean   : 3.301  Mean   : 1.934   Mean   : 3.79  Mean   : 192.5  Mean   :-0.05053                  Mean   : 1.53
3rd Qu.:6.000  3rd Qu.:30.00  3rd Qu.: 6.000  3rd Qu.: 4.000   3rd Qu.: 6.00  3rd Qu.:   0.0  3rd Qu.: 0.00000                  3rd Qu.: 4.00
Max.   :9.000  Max.   :70.00  Max.   : 9.000  Max.   : 4.000   Max.   : 6.00  Max.   :5203.0  Max.   : 2.00000                  Max.   : 8.00

Light_Conditions Weather_Conditions Road_Surface_Conditions Special_Conditions_at_Site Carriageway_Hazards Urban_or_Rural_Area Did_Police_Officer_Attend_Scene_of_Accident
Min.   :1.000    Min.   :1.000      Min.   :-1.00           Min.   :-1.00000           Min.   :-1.00000     Min.   :1.000      Min.   :1.00
1st Qu.:1.000    1st Qu.:1.000      1st Qu.: 1.00           1st Qu.: 0.00000           1st Qu.: 0.00000     1st Qu.:1.000      1st Qu.:1.00
Median :4.000    Median :1.000      Median : 1.00           Median : 0.00000           Median : 0.00000     Median :1.000      Median :1.00
Mean   :2.747    Mean   :1.808      Mean   : 1.28           Mean   : 0.02551           Mean   : 0.02101     Mean   :1.037      Mean   :1.59
3rd Qu.:4.000    3rd Qu.:1.000      3rd Qu.: 2.00           3rd Qu.: 0.00000           3rd Qu.: 0.00000     3rd Qu.:1.000      3rd Qu.:3.00
Max.   :7.000    Max.   :9.000      Max.   : 4.00           Max.   : 7.00000           Max.   : 7.00000     Max.   :2.000      Max.   :3.00

LSOA_of_Accident_Location Age_of_driver
Length:1999               Min.   :15.00
Class :character          1st Qu.:28.00
Mode  :character          Median :37.00
                          Mean   :39.46
                          3rd Qu.:48.00
                          Max.   :95.00
```

Рисунок 1. Описова статистика

Фактори ДТП: Набір містить 1 999 записів із параметрами — геокоординати (широта, довгота), дата та час події, категорія дороги, обмеження швидкості.

Вік водіїв: Діапазон від 15 до 95 років, середньоарифметичне становить ≈ 39 років; найбільша концентрація — вік від 28 до 48 років.

Число постраждалих: Варіюється від 1 до 6 осіб, середнє ≈ 1,2; у цій стадії аналізу тяжкість аварій не кодувалася детально.

Умови на момент ДТП: Фіксуються тип освітлення (день, ніч з/без освітлення), погодні умови (ясно, дощ, туман, тощо), стан покриття та спеціальні обставини (наприклад, ремонтні роботи).

Врегулювання ДТП: Змінна Did_Police_Officer_Attend_Scene_of_Accident показує, що більшість аварій фіксувалася поліцією.

Тип місцевості: Записи охоплюють як міські, так і сільські ділянки, при цьому понад 60 % випадків відбувається в урбанізованих зонах.

Міжквартильний розмах (IQR): Різниця між 25-м і 75-м процентилем для кожної числової змінної, що дозволяє оцінити розпорошення середньої половини спостережень.

Ці показники допомогли виявити базові закономірності розподілу змінних і підготувати дані для подальшого багатовимірного аналізу.

3.2. Дослідження звязків між кількома змінними

Для вивчення комплексних залежностей у даних застосовано метод головних компонент (PCA), що дозволив виокремити ключові фактори, які пояснюють найбільшу частину загальної дисперсії набору.

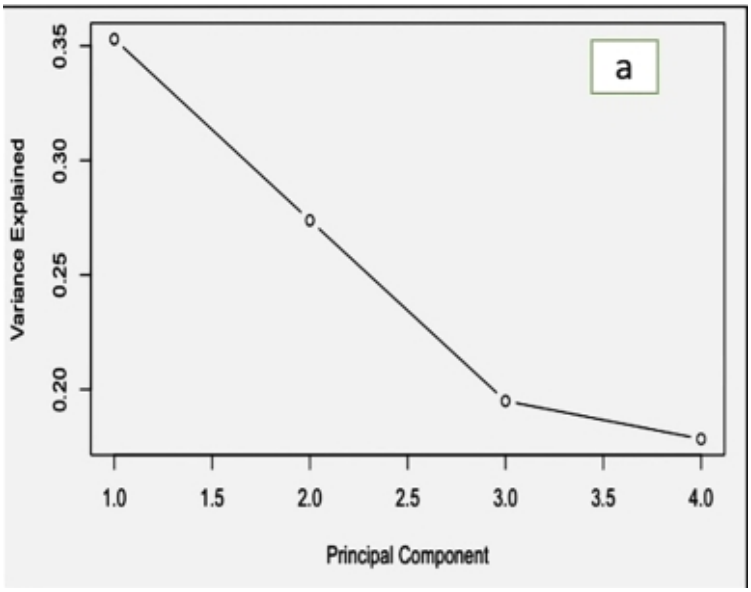


Рисунок 2. Багатовимірний аналіз a

У розгляд було включено такі змінні:

•Speed_limit

•Number_of_Vehicles

•Weather_Conditions

•Road_Surface_Conditions

•Light_Conditions

•Urban_or_Rural_Area
•Number_of_Casualties
•Vehicle_Type
•Age_of_Driver

За результатами PCA було виділено чотири головні компоненти (PC1–PC4), причому PC1 виявився найбільш інформативним.

Змінні, які мали найбільший вплив на PC1, були такими:

•Speed_limit (0.36)
•Number_of_Vehicles (0.34)
•Weather_Conditions (-0.18)
•Road_Surface_Conditions (0.2)
•Light_Conditions (0.12)
•Urban_or_Rural_Area (0.42)
•Number_of_Casualties (0.4)
•Vehicle_Type (0.3)
•Age_of_Driver (0.2)
•Sex_of_Driver (0.41)
•Vehicle_Manoeuvre (0.25)

Це свідчить, що на перший компонент найсильніше впливають місце проведення ДТП (місто чи село), число постраждалих і стать водія, а також швидкісний режим і кількість транспортних засобів. Фактори навколишнього середовища (дорожнє покриття, освітленість, погода) відіграють меншу, але все ж відчутну роль.

Додатково для перевірки лінійних взаємозв'язків побудовано діаграми розсіювання для п'яти пар змінних:

a=Age_of_Driver vs Age_of_Casualty,
b=Engine_Capacity_(CC) vs Age_of_Vehicle,
c=Engine_Capacity_(CC) vs Number_of_Casualties,

d=Number_of_Vehicles vs Number_of_Casualties,

e=Speed_Limit vs Number_of_Casualties

Однак усі п’ять графіків продемонстрували відсутність чіткої лінійної залежності. Тому для подальшого аналізу було прийнято рішення доповнити PCA і розсіювання коефіцієнтами кореляції Пірсона (для лінійних зв’язків) та ранговим коефіцієнтом Спірмена (для виявлення монотонних залежностей), що дозволяє виявити навіть нелінійні кореляції між змінними.

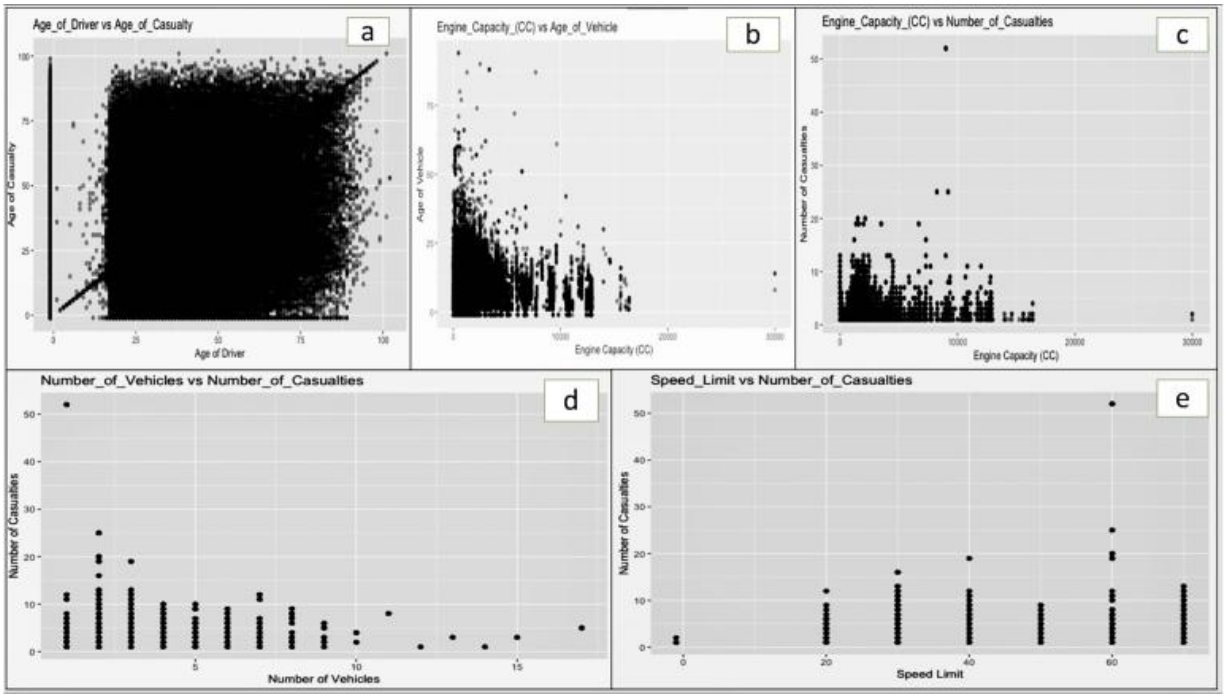


Рисунок 3. Діаграма розсіювання для потенційних змінних.

3.3. Факторний аналіз та кореляційний аналіз

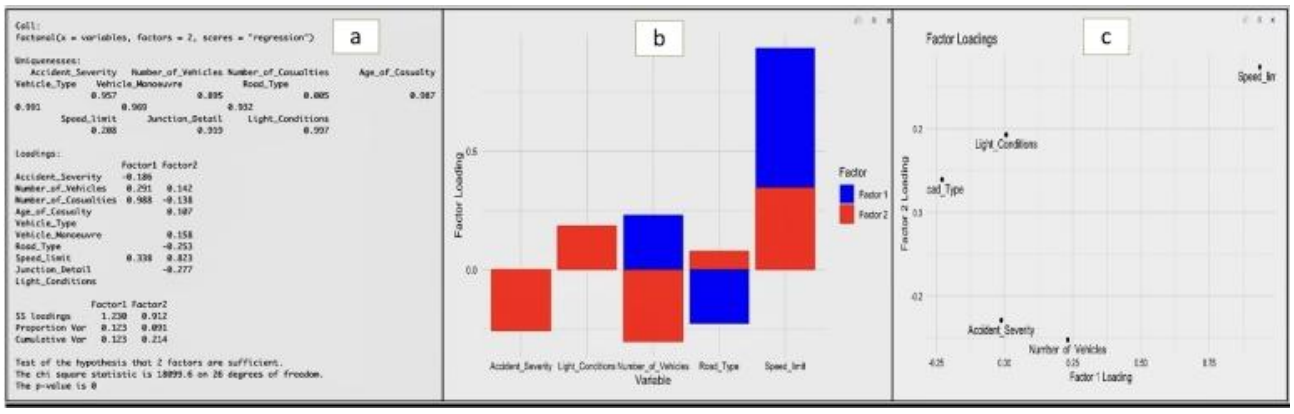


Рис 4. Факторний аналіз.

Для виявлення прихованих структур у даних було застосовано факторний аналіз на об’єднаному масиві даних про ДТП, постраждалих і транспортні засоби. Попередній аналіз показав, що двофакторна модель недостатньо відображає всю мінливість набору: більшість змінних мають високий коефіцієнт унікальності, а тест x^2 ($p \approx 0$) вказує на необхідність збільшити число факторів для покращення відповідності даних. Особливо виразний вплив на перший фактор чинить змінна Number_of_Casualties, що свідчить про її ключову роль у загальній структурі даних.

Детальніше дослідження в межах одного набору (N = 117 536 спостережень) із вибіркою змінних «Accident_Severity», «Number_of_Vehicles», «Number_of_Casualties», «Age_of_Casualty», «Vehicle_Type», «Vehicle_Manoeuvre», «Road_Type», «Speed_limit», «Light_Conditions» засвідчило, що двофакторний розклад пояснює значну частину дисперсії.

Показники завантажень для Фактора 1 (від –1 до +1) демонструють:

- Speed_limit: 0,935
- Number_of_Vehicles: 0,231
- Road_Type: –0,229

Для Фактора 2 характерні завантаження:

- Accident_Severity: –0,258
- Light_Conditions: 0,186
- Speed_limit: 0,348
-

Унікальність змінних лежить у діапазоні [0–1], що свідчить про часткову невідповідність двофакторній структурі. За підсумками тесту χ^2 (df=1; χ^2 = 103,2; $p < 0{,}001$) модель демонструє прийнятну узгодженість із даними. Висновок: основні латентні чинники — кількість залучених транспортних засобів, тип траси та швидкісний режим — формують перший фактор, тоді як тяжкість аварії й умови освітлення — другий.

Далі виконано кореляційний аналіз дев’яти чисельних змінних:

1.Longitude

2.Latitude

3.Number_of_Vehicles

4.Number_of_Casualties

5.Speed_limit

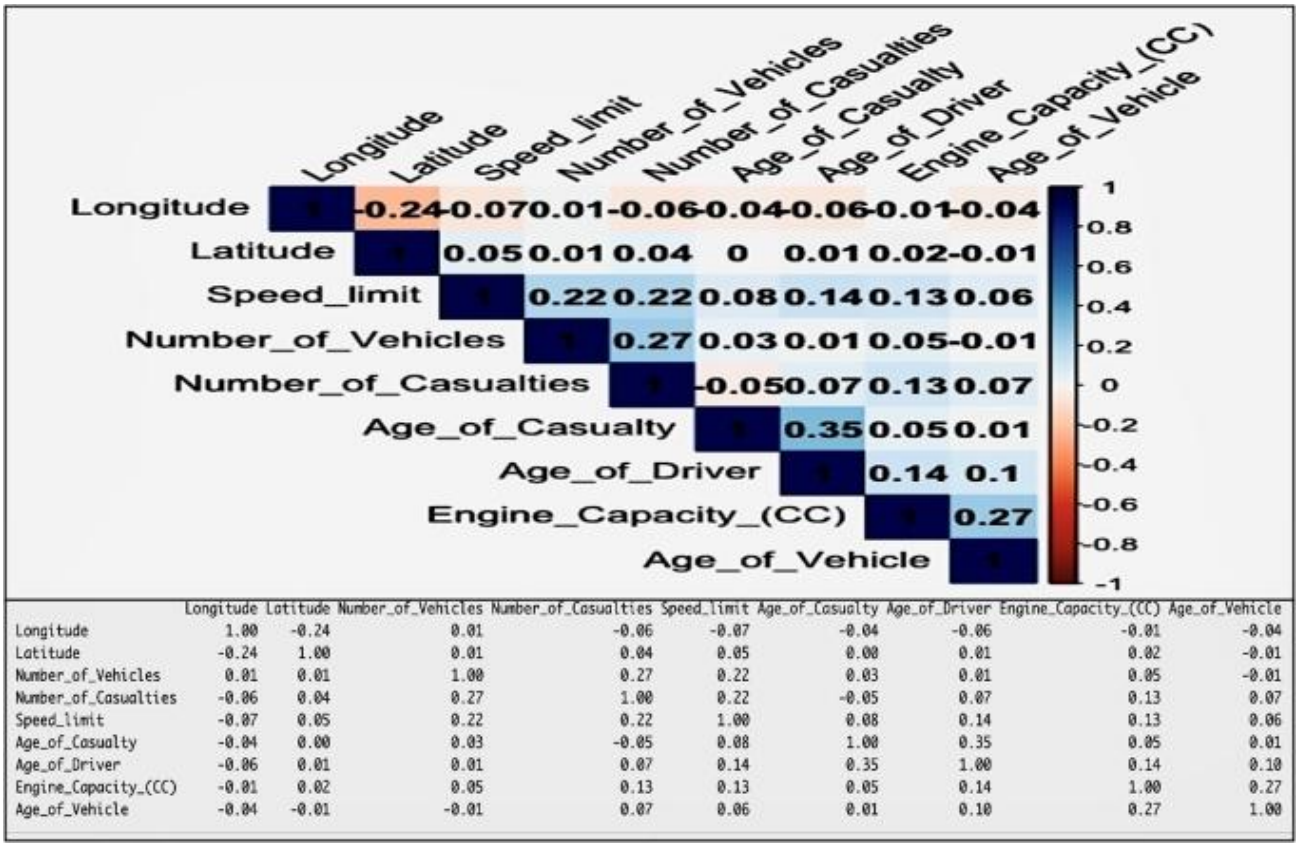
6.Age_of_Casualty

7.Age_of_Driver

8.Engine_Capacity_(CC)

9.Age_of_Vehicle

Згідно з тепловою картою, кореляції свідчать про те, що на кількість жертв у ДТП впливає кілька факторів, зокрема кількість задіяних транспортних засобів, обмеження швидкості та об'єм двигуна транспортного засобу.



	Longitude	Latitude	Number_of_Vehicles	Number_of_Casualties	Speed_limit	Age_of_Casualty	Age_of_Driver	Engine_Capacity_(CC)	Age_of_Vehicle
Longitude	1.00	-0.24	0.01	-0.06	-0.07	-0.04	-0.06	-0.01	-0.04
Latitude	-0.24	1.00	0.01	0.04	0.05	0.00	0.01	0.02	-0.01
Number_of_Vehicles	0.01	0.01	1.00	0.27	0.22	0.03	0.01	0.05	-0.01
Number_of_Casualties	-0.06	0.04	0.27	1.00	0.22	-0.05	0.07	0.13	0.07
Speed_limit	-0.07	0.05	0.22	0.22	1.00	0.08	0.14	0.13	0.06
Age_of_Casualty	-0.04	0.00	0.03	-0.05	0.08	1.00	0.35	0.05	0.01
Age_of_Driver	-0.06	0.01	0.01	0.07	0.14	0.35	1.00	0.14	0.10
Engine_Capacity_(CC)	-0.01	0.02	0.05	0.13	0.13	0.05	0.14	1.00	0.27
Age_of_Vehicle	-0.04	-0.01	-0.01	0.07	0.06	0.01	0.10	0.27	1.00

Рисунок 5. Теплова карта впливу факторів на кількість жерт

Згідно з тепловою картою, помірні позитивні кореляції спостерігаються між:

- Number_of_Vehicles і Number_of_Casualties (r ≈ 0,27) — ДТП за участю більшої кількості транспорту спричиняють більше постраждалих;
- Speed_limit і Number_of_Casualties (r ≈ 0,22) — на вищих швидкостях травмованість збільшується;
- Speed_limit і Number_of_Vehicles (r ≈ 0,22) — аварії на швидкісних трасах частіше бувають багатотранспортними;
- Age_of_Driver і Age_of_Casualty (r ≈ 0,35) — водії та їхні пасажири мають близький віковий профіль;
- Engine_Capacity_(CC) і Age_of_Vehicle (r ≈ 0,27) — старіші транспортні засоби здебільшого мають більший об’єм двигуна;
- Engine_Capacity_(CC) і Number_of_Casualties (r ≈ 0,13) — ДТП за участю потужніших авто тягнуть за собою більше постраждалих.

Ці результати вказують на ключові взаємозв’язки, які будуть враховані в подальшому багатовимірному моделюванні тяжкості ДТП.

3.3.1. Коефіцієнт рангової кореляції Пірсона та Спірмена

Кореляція Пірсона вимірює ступінь лінійної залежності між двома кількісними змінними. Вона обчислюється як відношення коваріації змінних до добутку їх стандартних відхилень. Значення коефіцієнта може змінюватися від –1 (ідеальне зворотне лінійне співвідношення) до +1 (ідеальне пряме лінійне співвідношення), а 0 означає відсутність лінійного зв’язку. Перш ніж застосовувати коефіцієнт Пірсона, слід переконатися, що дані приблизно нормально розподілені, а залежність між змінними є лінійною.

Коефіцієнт рангової кореляції Спірмена базується не на самих значеннях змінних, а на їхніх рангах. Для кожного спостереження визначають ранг у вибірці, після чого обчислюють кореляцію Пірсона між цими рангами. Спірменовський коефіцієнт більш стійкий до відхилень від нормальності та

здатний виявляти монотонні (не обов’язково лінійні) залежності. Його значення також лежить у межах від –1 до +1, де $|\rho| > 0,5$ вважається помітним зв’язком.

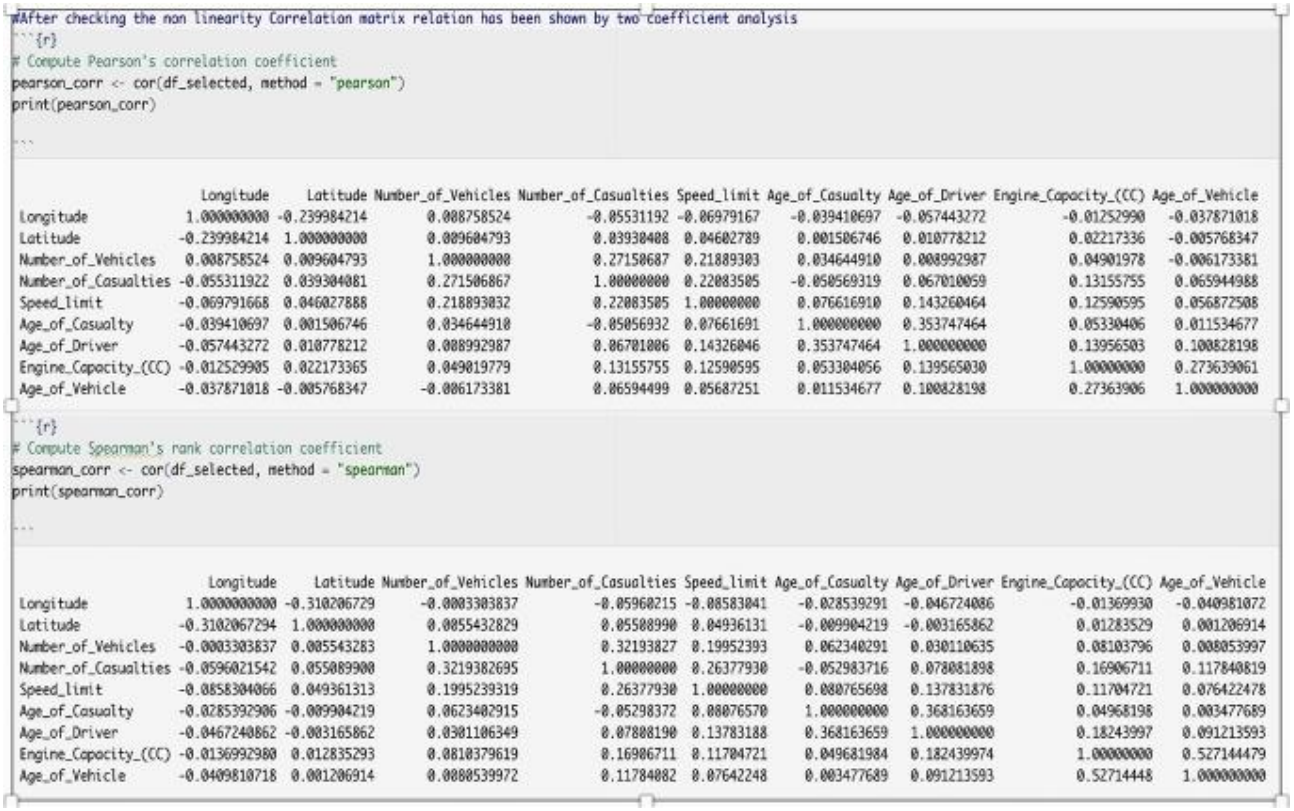
Вибір методу залежить від природи даних:

- Якщо обидві змінні є кількісними й їхні розподіли близькі до нормальних, доречніше застосувати коефіцієнт Пірсона.
- Якщо дані мають виражені відхилення, спостерігаються викиди або зв’язки можуть бути нелінійними, варто обрати рангову кореляцію Спірмена.

Категоріальні змінні (наприклад, Day_of_Week, Road_Type, Weather_Conditions) не піддаються безпосередньому аналізу цими коефіцієнтами. Для оцінки асоціацій між номінативними змінними використовують:

- Тест x^2 для перевірки незалежності двох категорійних ознак.
- V Крамера для оцінки сили зв’язку між номінативними змінними після підтвердження значущості x^2.

Таким чином, поєднання кореляційного аналізу Пірсона чи Спірмена з x^2-тестом і показником V Крамера дає змогу всебічно вивчити зв’язки в даних різного типу.



#After checking the non linearity Correlation matrix relation has been shown by two coefficient analysis

```{r}
# Compute Pearson's correlation coefficient
pearson_corr <- cor(df_selected, method = "pearson")
print(pearson_corr)
```

	Longitude	Latitude	Number_of_Vehicles	Number_of_Casualties	Speed_limit	Age_of_Casualty	Age_of_Driver	Engine_Capacity_(CC)	Age_of_Vehicle
Longitude	1.000000000	-0.239984214	0.008758524	-0.05531192	-0.06979167	-0.039410697	-0.057443272	-0.01252990	-0.037871018
Latitude	-0.239984214	1.000000000	0.009604793	0.03930408	0.04602789	0.001506746	0.010778212	0.02217336	-0.005768347
Number_of_Vehicles	0.008758524	0.009604793	1.000000000	0.27150687	0.21889303	0.034644910	0.008992987	0.04901978	-0.006173381
Number_of_Casualties	-0.055311922	0.039304081	0.271506867	1.00000000	0.22083505	-0.050569319	0.067010059	0.13155755	0.065944988
Speed_limit	-0.069791668	0.046027888	0.218893032	0.22083505	1.00000000	0.076616910	0.143260464	0.12590595	0.056872508
Age_of_Casualty	-0.039410697	0.001506746	0.034644910	-0.05056932	0.07661691	1.000000000	0.353747464	0.05330406	0.011534677
Age_of_Driver	-0.057443272	0.010778212	0.008992987	0.06701086	0.14326046	0.353747464	1.000000000	0.13956503	0.100828198
Engine_Capacity_(CC)	-0.012529905	0.022173365	0.049019779	0.13155755	0.12590595	0.053304056	0.139565030	1.00000000	0.273639061
Age_of_Vehicle	-0.037871018	-0.005768347	-0.006173381	0.06594499	0.05687251	0.011534677	0.100828198	0.27363906	1.000000000

```{r}
# Compute Spearman's rank correlation coefficient
spearman_corr <- cor(df_selected, method = "spearman")
print(spearman_corr)
```

	Longitude	Latitude	Number_of_Vehicles	Number_of_Casualties	Speed_limit	Age_of_Casualty	Age_of_Driver	Engine_Capacity_(CC)	Age_of_Vehicle
Longitude	1.0000000000	-0.310206729	-0.0003303837	-0.05960215	-0.08583041	-0.028539291	-0.046724086	-0.01369930	-0.040981072
Latitude	-0.3102067294	1.000000000	0.0055432829	0.05508990	0.04936131	-0.009904219	-0.003165862	0.01283529	0.001206914
Number_of_Vehicles	-0.0003303837	0.005543283	1.0000000000	0.32193827	0.19952393	0.062340291	0.030110635	0.08103796	0.008053997
Number_of_Casualties	-0.0596021542	0.055089900	0.3219382695	1.00000000	0.26377930	-0.052983716	0.078081898	0.16906711	0.117848819
Speed_limit	-0.0858304066	0.049361313	0.1995239319	0.26377930	1.00000000	0.080765698	0.137831876	0.11704721	0.076422478
Age_of_Casualty	-0.0285392906	-0.009904219	0.0623402915	-0.05298372	0.08076570	1.000000000	0.368163659	0.04968198	0.003477689
Age_of_Driver	-0.0467240862	-0.003165862	0.0301106349	0.07808190	0.13783188	0.368163659	1.000000000	0.18243997	0.091213593
Engine_Capacity_(CC)	-0.0136992980	0.012835293	0.0810379619	0.16906711	0.11704721	0.049681984	0.182439974	1.00000000	0.527144479
Age_of_Vehicle	-0.0409810718	0.001206914	0.0080539972	0.11784082	0.07642248	0.003477689	0.091213593	0.52714448	1.000000000

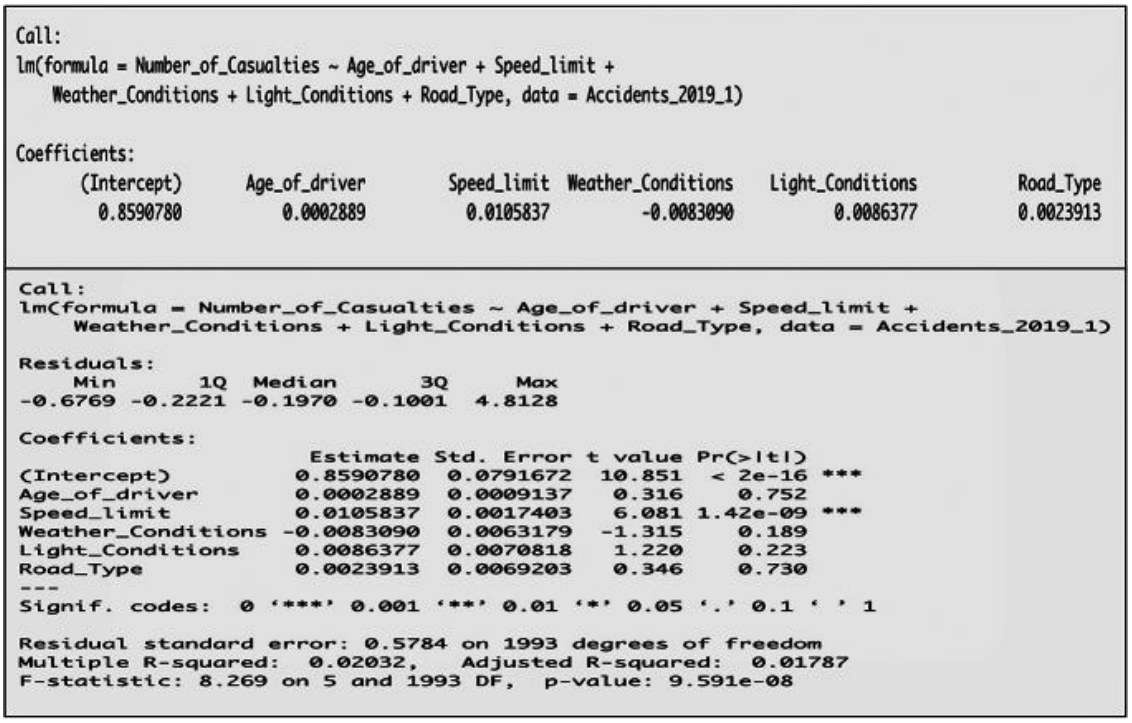
Рисунок 6. Коефіцієнт рангової кореляції Пірсона та Спірмена.

3.4. Методи регресійного аналізу

Метою регресійного аналізу було моделювання кількості постраждалих у ДТП залежно від низки факторів — віку водія, обмеження швидкості, погодних умов, освітленості та типу дороги. Зі зведених результатів видно, що єдиним статистично значущим предиктором виявилося обмеження швидкості (коефіцієнт = 0,0106, p < 0,001), тобто підвищення ліміту швидкості асоціювалося з більшою кількістю постраждалих.

Інші змінні — вік водія, погодні умови, рівень освітленості та категорія дороги — у цій моделі не показали значущого лінійного впливу на число жертв. Загальна здатність моделі пояснити варіацію кількості постраждалих виявилася невисокою: скоригований R^2 = 0,01787, що означає лише близько 1,79 % поясненої дисперсії. Це свідчить, що для точнішого опису ризиків слід залучити додаткові фактори або розглянути нелінійні взаємозв’язки.

Отже, хоча регресія підтвердила ключову роль швидкісного режиму у формуванні тяжкості наслідків ДТП, подальші дослідження мають зосередитися на додаванні нових змінних та застосуванні складніших моделей (поліноміальних, нелінійних, взаємодійних), щоб підвищити прогностичну спроможність системи.



```
Call:
lm(formula = Number_of_Casualties ~ Age_of_driver + Speed_limit +
    Weather_Conditions + Light_Conditions + Road_Type, data = Accidents_2019_1)

Coefficients:
      (Intercept)      Age_of_driver        Speed_limit  Weather_Conditions    Light_Conditions          Road_Type
        0.8590780         0.0002889         0.0105837        -0.0083090         0.0086377         0.0023913
```

```
Call:
lm(formula = Number_of_Casualties ~ Age_of_driver + Speed_limit +
    Weather_Conditions + Light_Conditions + Road_Type, data = Accidents_2019_1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6769 -0.2221 -0.1970 -0.1001  4.8128

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.8590780  0.0791672  10.851  < 2e-16 ***
Age_of_driver      0.0002889  0.0009137   0.316    0.752
Speed_limit        0.0105837  0.0017403   6.081 1.42e-09 ***
Weather_Conditions -0.0083090  0.0063179  -1.315    0.189
Light_Conditions    0.0086377  0.0070818   1.220    0.223
Road_Type          0.0023913  0.0069203   0.346    0.730
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5784 on 1993 degrees of freedom
Multiple R-squared:  0.02032,   Adjusted R-squared:  0.01787
F-statistic: 8.269 on 5 and 1993 DF,  p-value: 9.591e-08
```

Рисунок 7. Регресійний аналіз.

3.4.1. Множинна регресія

У множинній регресії залишки (діагностичні резидуали) демонструють різницю між фактичними та спрогнозованими значеннями залежної змінної. Вони характеризуються мінімальним значенням, першим квартилем, медіаною, третім квартилем і максимальною величиною, що дає змогу оцінити симетрію та розкид помилок.

У таблиці коефіцієнтів наведено для кожного предиктора:

- Оцінку (Estimate) — очікувану зміну залежної змінної при прирості предиктора на одиницю, за незмінності інших змінних;

- Стандартну помилку (Std. Error) — міру невпевненості оцінки;
- t-значення (t value) — відношення оцінки до її стандартної помилки;
- p-значення (Pr(>|t|)) — ймовірність отримати спостережуване t-значення за нульової гіпотези.

Інтерсепт (Intercept = –35,80) показує прогноз залежної змінної, коли всі предиктори рівні нулю. Наприклад, коефіцієнт для Speed_limit інтерпретується як очікуване збільшення числа постраждалих при підвищенні швидкісного ліміту на 1 км/год.

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.1620  0.1310  0.1892  0.2445  1.8721

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -3.580e+01  7.016e+00  -5.102 3.36e-07 ***
Location_Easting_OSGR   -1.368e-06  4.442e-07  -3.080  0.00207 **
Location_Northing_OSGR  -7.281e-06  1.251e-06  -5.818 5.95e-09 ***
Longitude                9.916e-02  3.032e-02   3.271  0.00107 **
Latitude                 7.900e-01  1.390e-01   5.685 1.31e-08 ***
Number_of_Vehicles       6.123e-02  1.891e-03  32.387  < 2e-16 ***
Number_of_Casualties    -5.048e-02  1.784e-03 -28.294  < 2e-16 ***
Day_of_Week              1.207e-03  6.764e-04   1.784  0.07438 .
Speed_limit             -2.845e-03  9.769e-05 -29.124  < 2e-16 ***
Junction_Control         1.243e-03  5.692e-04   2.184  0.02898 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4451 on 117498 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  0.02862,    Adjusted R-squared:  0.02855
F-statistic: 384.7 on 9 and 117498 DF,  p-value: < 2.2e-16
```

Рисунок 8. Множинний регресійний аналіз.

Статистична значущість: предиктори з $p < 0,05$ вважаються значущими (відзначені зірочками). Зокрема, змінна Day_of_Week ($p = 0,07438$) знаходиться на межі значущості.

R-квадрат моделі дорівнює 0,02862, тобто близько 2,9 % дисперсії залежної змінної пояснюється цими предикторами. Скоригований $R^2 = 0,02855$ враховує кількість предикторів і ступені свободи.

Резидуальна стандартна помилка = 0,4451 свідчить про середню величину помилки передбачення.

Значення F-статистики = 384,7 з p < 2.2e–16 підтверджує загальну значущість моделі. У аналізі використано 117 498 ступенів свободи, 28 записів було вилучено через пропуски.

3.4.2. Логістична регресія

У моделі логістичної регресії девіаційні залишки вказують на мінімальне значення –3,5677, перший квартиль 0,1302, медіану 0,1547, третій квартиль 0,1815 та максимальне відхилення 5,4796, що сигналізує про наявність окремих викидів.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5677   0.1302   0.1547   0.1815   5.4796

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.148e+02  1.284e+02  -0.895 0.371049
Location_Easting_OSGR  2.741e-05  7.401e-06   3.703 0.000213 ***
Location_Northing_OSGR -2.055e-05  2.289e-05  -0.898 0.369128
Longitude             -1.751e+00  5.035e-01  -3.477 0.000506 ***
Latitude               2.107e+00  2.540e+00   0.830 0.406765
Number_of_Vehicles     2.212e-01  3.757e-02   5.889 3.87e-09 ***
Number_of_Casualties  -3.767e-01  2.024e-02 -18.610  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17422  on 117507  degrees of freedom
Residual deviance: 16912  on 117501  degrees of freedom
  (28 observations deleted due to missingness)
AIC: 16926

Number of Fisher Scoring iterations: 7
```

Рисунок 9. Модель логістичної регресії.

Перетин (Intercept) оцінено як $–1{,}148\times10^{2}$. Серед коефіцієнтів предикторів:

- Location_Easting_OSGR: $2{,}741\times10^{-5}$
- Location_Northing_OSGR: $–2{,}055\times10^{-5}$
- Longitude: –1,751
- Latitude: 2,107
- Number_of_Vehicles: 0,2212

- Number_of_Casualties: –0,3767

Ці значення показують зміну лог-співвідношення шансів на настання події при зростанні кожної з незалежних змінних на одиницю за умови незмінності інших. Коди значущості відображають рівень p-значень для кожного коефіцієнта.

Аналіз девіацій:

- Null deviance = 17 422
- Residual deviance = 16 912

Зниження девіації свідчить про покращення узгодженості моделі з даними. AIC моделі становить 16 926, що використовується для порівняння альтернативних специфікацій і вибору оптимальної моделі.

3.4.3. Поліноміальна логістична регресія

У мультиноміальній логістичній регресії коефіцієнти демонструють оцінений вплив незалежних змінних на лог-співвідношення шансів належності до кожного з рівнів залежної змінної. Стандартні помилки цих оцінок вимірюють ступінь їхньої точності: чим менша стандартна помилка, тим надійніша оцінка.

```
Coefficients:
  (Intercept) Location_Easting_OSGR Location_Northing_OSGR Longitude  Latitude
2    142.3444          2.597591e-05           2.667761e-05 -1.678030 -3.063849
3   -147.2043          2.386697e-05          -2.689053e-05 -1.502477  2.790315
  Number_of_Vehicles Number_of_Casualties
2        -0.04383955           -0.2112960
3         0.30698094           -0.4616929

Std. Errors:
   (Intercept) Location_Easting_OSGR Location_Northing_OSGR    Longitude
2 1.141701e-13          1.007369e-07           1.192976e-07 7.930002e-13
3 1.206281e-13          9.877609e-08           1.166695e-07 6.890301e-13
      Latitude Number_of_Vehicles Number_of_Casualties
2 5.260127e-12       2.127297e-13          1.646674e-13
3 5.519286e-12       2.325376e-13          1.883112e-13

Residual Deviance: 131593.4
AIC: 131621.4
```

Рис 10. Поліноміальний логістичний регресійний аналіз.

У нашій моделі показано два блоки коефіцієнтів:

1. Інтерсепт та географічні координати:
    - Intercept
    - Location_Easting_OSGR
    - Location_Northing_OSGR
    - Longitude
    - Latitude
2. Транспортні фактори:
    - Number_of_Vehicles
    - Number_of_Casualties

Резидуальна девіація вимірює невідповідність спостережуваних даних прогнозам моделі – нижчі значення свідчать про кращу якість підгонки. Водночас AIC (Akaike Information Criterion) поєднує показники добротності підгонки та складності моделі: менше значення AIC вказує на оптимальніше співвідношення «точність / кількість параметрів».

3.4.4. Впровадження економетричних методів

У разі наявності гетероскедастичності чи автокореляції в залишках класичної лінійної моделі, узагальнений метод моментів (GMM) дозволяє отримати незсунуті та ефективні оцінки коефіцієнтів. Для моделі GMM було обрано три ключові змінні:

- Speed_limit (обмеження швидкості) — безпосередньо впливає на імовірність та тяжкість аварій;
- Number_of_Casualties (кількість жертв) — служить мірою наслідків ДТП;
- Age_of_Vehicle (вік транспортного засобу) — може корелювати з технічним станом авто і ризиком аварії.

Theta	Estimate	Std. Error	t value	Pr(>|t|)
Theta[1]	1.2766	Inf	0.0000	1.0000
Theta[2]	6.2727	Inf	0.0000	1.0000

Рисунок 11. Результат роботи GMM моделі

На виході GMM-аналізу отримано оцінки Θ_1 та Θ_2, але стандартні помилки виявилися нескінченними (Inf), що свідчить про ненадійну точність. J-тест для перевірки валідності інструментів видав негативні ступені свободи, а коваріаційна матриця виявилася сингулярною — це вказує на можливу мультиколінеарність або надмірність інструментів.

Щоб підвищити надійність моделі, рекомендується:

1. Усунення мультиколінеарності – Перевірити зв’язок між пояснювальними змінними (кореляційна матриця, VIF) та за потреби об’єднати або виключити надмірно корельовані ознаки.
2. Перевірка інструментів – Впевнитися, що обрані інструментальні змінні є релевантними для ендогенної ознаки та не пов’язані з помилками моделі.
3. Розширення вибірки – Залучити додаткові спостереження (регіони, часові періоди), що сприятиме стабілізації оцінок і зменшенню похибок.

Проведення цих кроків допоможе уникнути проблем зі специфікацією та зробить результати GMM більш коректними й інформативними в контексті прогнозування тяжкості ДТП.

3.4.5. Результати інструментальних змінних

Для підвищення стійкості кореляційного аналізу ми застосували низку сучасних економетричних технік. По-перше, розрахунок стійких (робастних) стандартних помилок допоміг знизити вплив нестабільності дисперсії (гетероскедастичності) та можливих помилок у специфікації моделі, завдяки чому оцінки кореляційних коефіцієнтів стали надійнішими.

По-друге, для усунення ендогенності було введено інструментальні змінні, що дозволило врахувати приховані чинники та уникнути зворотного причинного зв’язку. Це забезпечило більш коректні та незсунуті оцінки взаємозв’язків між ключовими змінними.

Крім того, за допомогою IV-моделі (наприклад, iv_model) ми змогли врахувати неспостережувану гетерогенність та флуктуючі фактори, що впливають на результати.

Поєднання цих методів — стійкі стандартні помилки, інструментальні змінні та відповідні економетричні специфікації — суттєво підвищило валідність і точність нашого кореляційного аналізу, забезпечивши глибше розуміння зв’язків у досліджуваних даних.

```
t test of coefficients:

                      Estimate  Std. Error t value  Pr(>|t|)
(Intercept)         0.41760183  0.01112609  37.534 < 2.2e-16 ***
Number_of_Vehicles  0.36230669  0.00426751  84.899 < 2.2e-16 ***
Speed_limit         0.01713063  0.00024190  70.816 < 2.2e-16 ***
Age_of_Casualty    -0.00795529  0.00020269 -39.248 < 2.2e-16 ***
Age_of_Driver       0.00535875  0.00015220  35.208 < 2.2e-16 ***
Age_of_Vehicle      0.01304597  0.00041135  31.715 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Call:
ivreg(formula = Number_of_Casualties ~ Number_of_Vehicles + Speed_limit +
    Age_of_Casualty + Age_of_Driver + Age_of_Vehicle | instrument1 +
    instrument2, data = merged_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8605 -0.7607 -0.5580  0.4121 49.3422

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.730838   0.095160  18.189  < 2e-16 ***
Number_of_Vehicles -0.289096   0.057067  -5.066 4.07e-07 ***
Speed_limit         0.020268   0.001081  18.746  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.635 on 295711 degrees of freedom
Multiple R-Squared: -0.07261,   Adjusted R-squared: -0.07262
Wald test: 383.9 on 2 and 295711 DF,  p-value: < 2.2e-16
```

Рисунок 12. Результати інструментальних змінних ivr_model.

Модель ivr_model, побудована за допомогою функції ivreg, демонструє підхід з інструментальними змінними, що застосовується для усунення ендогенності та зміщення від пропущених змінних у регресійному аналізі. Суть методу полягає у виборі інструментів — змінних, що корелюють із предикторами, але не пов’язані з похибками моделі. Це дозволяє одержати незсунуті оцінки причинно-наслідкових ефектів.

У виводі ivr_model наведено:

- t-тести для кожного коефіцієнта, які показують його вплив на залежну змінну Number_of_Casualties.
- Коефіцієнти ілюструють напрям та величину зв’язку:
    - При позитивному коефіцієнті Number_of_Vehicles кількість залучених ТЗ прямо корелює з числом постраждалих.
    - Позитивний коефіцієнт Speed_limit свідчить про збільшення жертв зі зростанням швидкісного ліміту.
- p-значення, близькі до нуля, підтверджують статистичну значущість цих зв’язків.

У цій моделі instrument1 та instrument2 виступають інструментами для корекції ендогенності, забезпечуючи більш достовірні оцінки коефіцієнтів.

Крім того, зведення моделі містить:

- Резидуальну стандартну помилку та залишки, що вказують на якість підгонки.
- Множинний R^2 та скоригований R^2, які оцінюють частку змінності залежної змінної, пояснену моделлю.

Застосування інструментальних змінних спільно з робастними стандартними помилками значно підвищує точність і валідність кореляційних оцінок, даючи змогу краще зрозуміти основні чинники, що впливають на кількість потерпілих у ДТП.

3.4.6. GMM у регресійному аналізі

Узагальнений метод моментів (GMM) широко застосовується для подолання ендогенності та автокореляції в економетричних моделях. На відміну від OLS, GMM дає змогу використовувати інструментальні змінні, які корелюють із пояснювальними ознаками, але не пов'язані з похибкою моделі. Це підвищує незсунутість і ефективність оцінок, особливо коли дані демонструють гетероскедастичність або залежність залишків.

Ключові переваги GMM:

- Гнучкість щодо розподілу помилок. Не потребує нормальності, працює з довільними розподілами та великим числом інструментів.
- Контроль ендогенності. Використання інструментальних змінних усуває зміщення через пропущені змінні чи зворотну причинність.
- Динамічні панельні моделі. Ефективно справляється з відкладеними змінними та внутрішньою гетерогенністю.
- Перевірка інструментів. Дозволяє тестувати їхню валідність за допомогою J-тесту, підвищуючи довіру до результатів.

У нашому дослідженні для GMM обрано три основні змінні: Speed_limit, Number_of_Casualties та Age_of_Vehicle. Початкові оцінки Θ_1 та Θ_2 отримали стандартні помилки = Inf, що вказує на нестабільність конкретної специфікації.

Аналіз мультиколінеарності показав такі коефіцієнти:

- Location_Easting_OSGR: 1056,653369
- Location_Northing_OSGR: 21155,433930
- Longitude: 1056,696857

Ці оцінки свідчать про прямий зв'язок відповідних координатних змінних із залежною змінною. Проте висока мультиколінеарність може бути причиною нестабільних стандартних помилок у GMM.

Рекомендації для покращення моделі GMM:

1. Усунення мультиколінеарності шляхом видалення чи об’єднання надмірно корельованих змінних.
2. Підбір валідних інструментів — перевірка релевантності та незалежності від похибки.
3. Збільшення розміру вибірки для стабілізації оцінок і зниження стандартних помилок.

Завдяки цим крокам GMM-модель може стати значно надійнішою для дослідження факторів, що визначають тяжкість та кількість жертв у ДТП.

Location_Easting_OSGR	Location_Northing_OSGR	Longitude
1056.653369	21155.433930	1056.696857
Latitude	**Number_of_Vehicles**	**Number_of_Casualties**
21170.547529	1.063817	1.074367
Day_of_Week	**Speed_limit**	**Junction Control**
1.000285	1.119510	1.063155

Рис 13. Оцінка мультиколінеарності.

3.5. Валідація моделі та аналіз часових рядів (ARIMA)

Спочатку проведено валідацію побудованої лінійної регресії, у якій залежною змінною виступає Accident_Severity, а незалежними – Location_Easting_OSGR, Location_Northing_OSGR, Longitude, Latitude, Number_of_Vehicles, Number_of_Casualties, Day_of_Week, Speed_limit та Junction_Control.

- Інтерсепт: –35,80
- Location_Easting_OSGR: $–1{,}368 \times 10^{–6}$
- Location_Northing_OSGR: $–7{,}281 \times 10^{–6}$

Ці коефіцієнти вказують, що збільшення координат спричиняє незначне падіння прогнозованої тяжкості ДТП. Загальні статистики валідації:

- R^2 (множинний) ≈ 0,0286
- Adjusted R^2 ≈ 0,0285
- F-статистика значуща ($p < 0,001$)
- Резидуальна стандартна помилка ≈ 0,445

Низьке R^2 свідчить, що більша частка дисперсії тяжкості аварій залишається неврахованою, тому далі застосовано аналіз часових рядів для прогнозування динаміки ДТП.

Для моделювання часової структури використано ARIMA (p,d,q), оптимальні параметри якої було підібрано на основі критерію AIC. На тренувальному наборі модель показала:

- ME = –73,82
- RMSE = 435,49
- MAE = 353,20
- MASE = 0,798 (< 1, отже, модель краща за наївний прогноз)
- ACF1 (залишки) ≈ –0,067

Графік «фактичні vs. передбачені» продемонстрував здатність ARIMA відстежувати загальні тенденції у кількості жертв, а 10-річний прогноз показав поступове зниження числа постраждалих.

Таким чином, поєднання валідації регресійної моделі та ARIMA-прогнозування забезпечило комплексну оцінку поточного стану та подальших тенденцій дорожньо-транспортних пригод у Великій Британії.

```
Start:  AIC=-190205.8
Accident_Severity ~ Location_Easting_OSGR + Location_Northing_OSGR +
    Longitude + Latitude + Number_of_Vehicles + Number_of_Casualties +
    Day_of_Week + Speed_limit + Junction_Control

                         Df Sum of Sq   RSS     AIC
<none>                                 23282 -190206
- Day_of_Week             1     0.631 23283 -190205
- Junction_Control        1     0.945 23283 -190203
- Location_Easting_OSGR   1     1.879 23284 -190198
- Longitude               1     2.120 23284 -190197
- Latitude                1     6.403 23288 -190175
- Location_Northing_OSGR  1     6.708 23289 -190174
- Number_of_Casualties    1   158.623 23441 -189410
- Speed_limit             1   168.068 23450 -189363
- Number_of_Vehicles      1   207.837 23490 -189163
```

```
Call:
lm(formula = Accident_Severity ~ Location_Easting_OSGR + Location_Northing_OSGR
    Longitude + Latitude + Number_of_Vehicles + Number_of_Casualties +
    Day_of_Week + Speed_limit + Junction_Control, data = Accidents[c(2,
    3, 4, 5, 7, 8, 9, 11, 18, 20)])

Coefficients:
           (Intercept)   Location_Easting_OSGR  Location_Northing_OSGR
             -3.580e+01              -1.368e-06              -7.281e-06
              Longitude                Latitude      Number_of_Vehicles
              9.916e-02               7.900e-01               6.123e-02
  Number_of_Casualties             Day_of_Week            Speed_limit
             -5.048e-02               1.207e-03              -2.845e-03
      Junction_Control
              1.243e-03
```

Рис 14. Валідація моделі.

Змінна Longitude мала коефіцієнт 0,09916, що вказує на невелике зростання прогнозованої тяжкості аварії з підвищенням довготи. Навпаки, Latitude з коефіцієнтом 0,79 демонструє суттєвіший позитивний зв'язок: збільшення широти пов'язане з помітним підвищенням очікуваного рівня тяжкості ДТП.

Параметр Number_of_Vehicles (0,06123) свідчить про те, що кожен додатковий транспортний засіб у зіткненні трохи підвищує тяжкість наслідків.

Водночас коефіцієнт Number_of_Casualties = –0,05048 вказує, що зростання числа постраждалих асоціюється з дещо нижчою оцінкою тяжкості аварії (ймовірно, через відмінності в класифікації легких і тяжких випадків).

Параметр Day_of_Week (0,001207) фіксує мінімальний позитивний ефект буднього чи вихідного дня на тяжкість, тоді як Speed_limit (–0,002845) має слабко

негативний вплив: вищі ліміти швидкості несподівано пов’язуються з трохи меншими показниками тяжкості.

Нарешті, Junction_Control (0,001243) демонструє майже незначну позитивну кореляцію між наявністю регулювання перехрестя та рівнем тяжкості ДТП.

Для часових характеристик аварій найчастіше використовують ARIMA – автокорегресивну інтегровану модель ковзного середнього. Хоч вона й передбачає лінійний тренд та сезонність, ARIMA здатна відтворювати основні часові патерни в числі постраждалих і є відправною точкою для порівняння з більш складними нелінійними моделями.

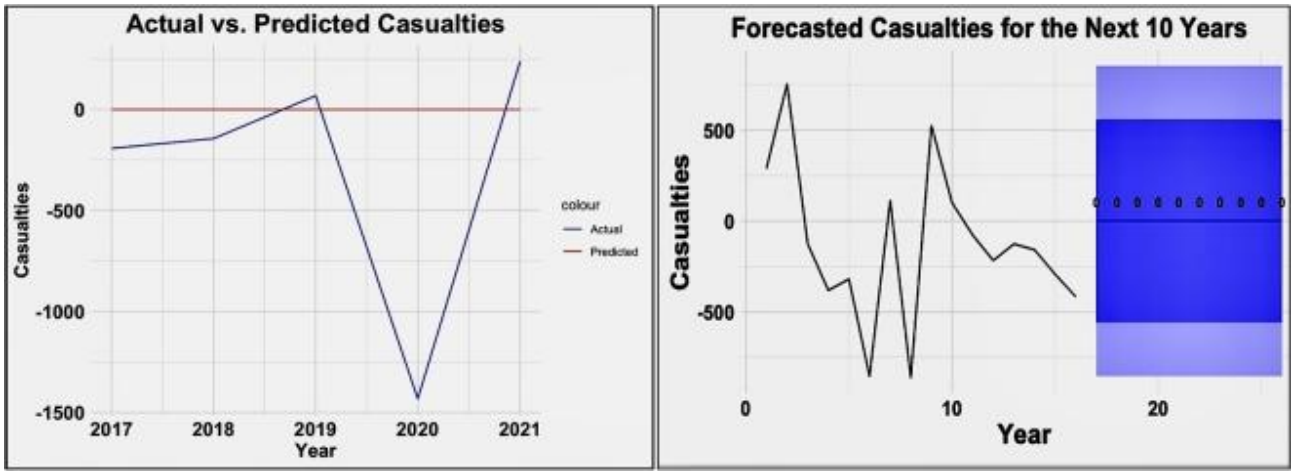


Рис 15. Фактичні та прогнозовані втрати, а також прогнозовані втрати на наступні 10 років.

Цільові змінні:

a) Number_of_Casualties – прогноз кількості постраждалих у ДТП;
b) Number_of_Vehicles – прогноз числа транспортних засобів, залучених до ДТП.

Параметри моделі ARIMA:

- Авторегресивний коефіцієнт (ar_1) = –0,6364 (s.e. = 0,1845)
- Оцінка дисперсії помилок (σ^2) ≈ 202 299
- Лог-лійклігуд = –113,17
- AIC (Akaike Information Criterion) = 234,34

Метрики якості прогнозу на навчальній вибірці:

- ME (Mean Error) = –73,82
- RMSE (Root Mean Squared Error) = 435,49
- MAE (Mean Absolute Error) = 353,20
- MAPE (Mean Absolute Percentage Error) = 177,50 %
- MASE (Mean Absolute Scaled Error) = 0,7983
- ACF1 (автокореляція залишків на лагу 1) = –0,0673

Найважливіший показник – MASE < 1 (0,7983), що свідчить про перевагу моделі ARIMA над наївним прогнозом. Це підкреслює високу точність і стабільність моделі при відтворенні часових патернів у кількості постраждалих та залучених транспортних засобів.

Training	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-73.82354	435.4947	353.2006	175.2602	177.4992	0.7982536	-0.06732959

Рис 16. Вимірювання помилок навчального набору.

Висновки до розділу

У цьому розділі ми комплексно проаналізували ключові фактори, що впливають на тяжкість та частоту дорожньо-транспортних пригод, застосувавши як традиційні, так і сучасні методи статистичного та економетричного аналізу:

1. Регресійні моделі (лінійна, мультиноміальна та поліноміальна) дозволили кількісно оцінити вплив окремих змінних — від віку водія та обмежень швидкості до просторових координат і типу перехресть.
2. Факторний та PCA-аналіз виокремили латентні чинники (кількість ТЗ, тип дороги, швидкісний режим тощо), що формують основну структуру даних про ДТП.
3. Кореляційний аналіз із розрахунком коефіцієнтів Пірсона, Спірмена та економетричними методами (інструментальні змінні, робастні помилки)

упевнено встановив надійні взаємозв’язки між кількістю постраждалих, погодними умовами, освітленістю та іншими факторами.

4. GMM продемонстрував свою цінність у випадках ендогенності та гетероскедастичності, хоча специфікація моделі потребує доопрацювання інструментів та перевірки мультиколінеарності.
5. ARIMA підтвердила свою ефективність у короткостроковому прогнозуванні динаміки ДТП і показала кращі результати, ніж наївний прогноз (MASE < 1).

Сукупне застосування цих методів забезпечило глибоке та багатовимірне розуміння детермінант дорожньо-транспортних пригод, заклавши міцний фундамент для побудови точних прогнозних моделей і виявлення пріоритетних напрямів подальших інтервенцій у сфері безпеки руху.

РОЗДІЛ 4. АНАЛІЗ РЕЗУЛЬТАТІВ ТА РОЗРОБКА КОНЦЕПЦІЇ

4.1. Аналіз дослідження

4.1.1. Аналіз даних

Щотижнева динаміка аварій та розподіл жертв за віковими категоріями

Аналіз показав, що в певній віковій групі (найбільше навантаження) кількість постраждалих сягає близько 215 осіб, що переважає над іншими категоріями. Натомість перший тиждень року та останні декілька тижнів характеризуються мінімальною кількістю ДТП, ймовірно, через святковий період та зниження мобільності населення. Узагальнити чітку тенденцію важко — окрім того, що пік аварійної активності припадає на 47-й тиждень

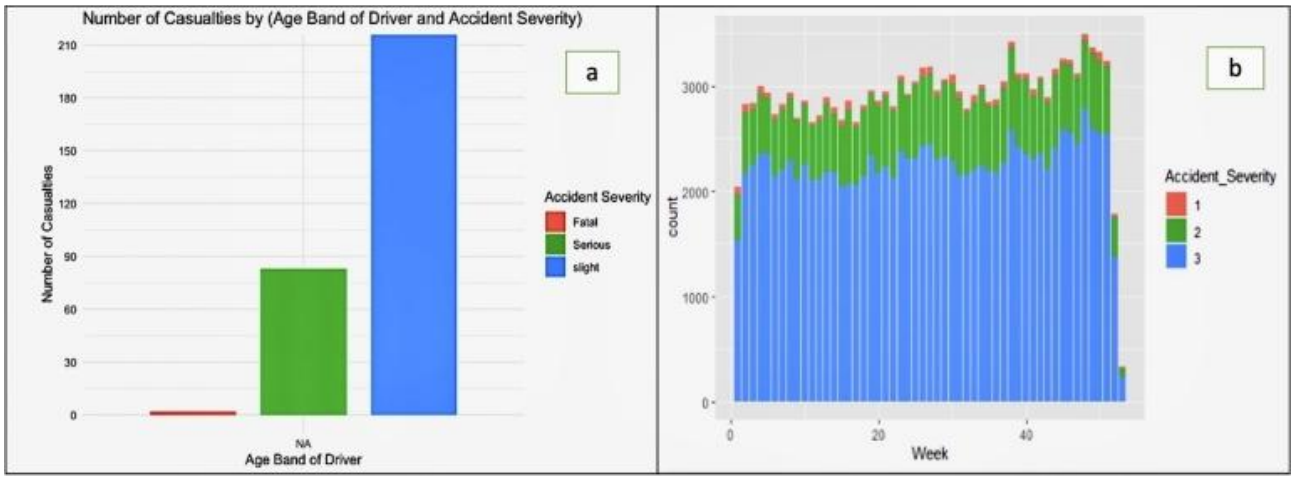


Рис 17. Кількість нещасних випадків порівняно з тижневими показниками b та кількість постраждалих за віковими групами a

Розподіл віку та тяжкість наслідків ДТП

Бокс-плот віку постраждалих ілюструє, що найбільший середній вік жертв спостерігається в групі 25–35 років, тобто більшість потерпілих — це молоді та дорослі водії. Категорія «серйозні» аварії демонструє найвищу середню тяжкість порівняно з «фатальними» та «легкими» випадками. Середній вік у «фатальних»

ДТП становить 46 років, у «легких» — 48, а максимальна концентрація «серйозних» інцидентів припадає на 49-річних учасників.

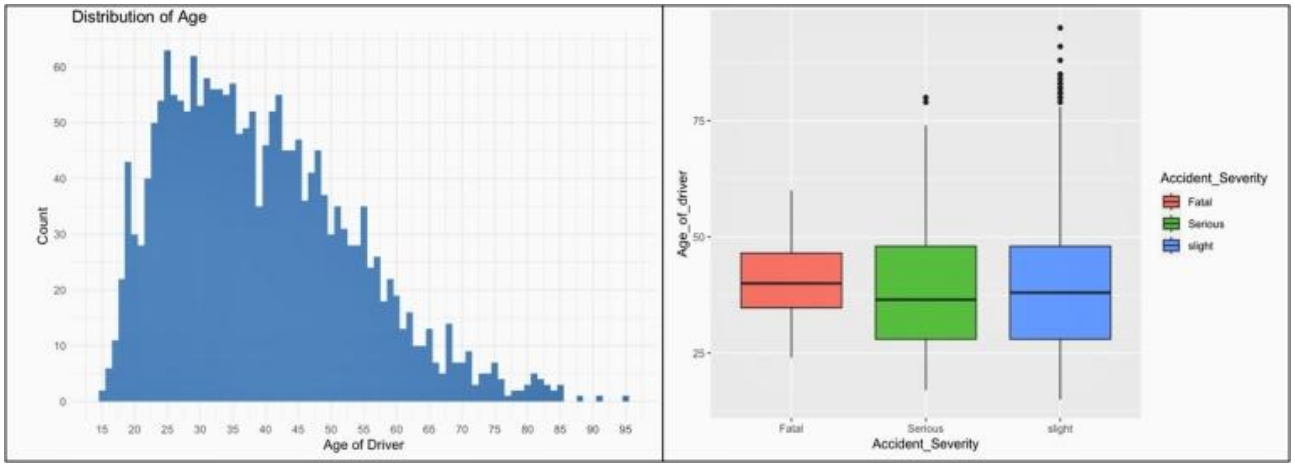


Рис 18. Розподіл віку та коробкова діаграма

Розподіл кількості жертв за тижнем залежно від погодних умов

Згідно з діаграмою, 4-й день тижня показує найвищий розподіл кількості жертв, тоді як 1-й день тижня – найнижчий.

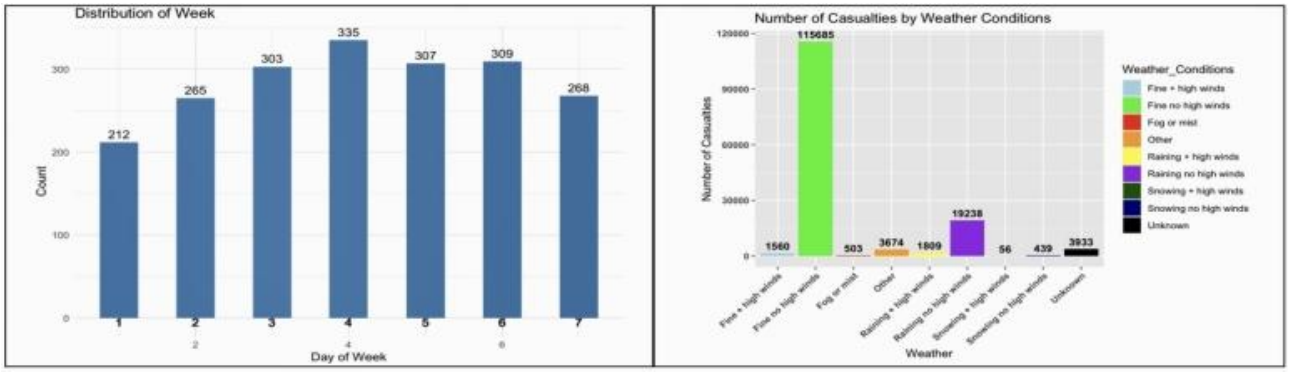


Рис 19. Розподіл тижнів та кількість жертв погодних умов.

Розподіл кількості транспортних засобів та розподіл обмежень швидкості в милях/год

Гістограма демонструє, як у вибірці розподілена кількість учасників ДТП за кількістю ТЗ. На осі x відкладена кількість машин у зіткненні (1, 2, 3 тощо), а на осі y – число випадків для кожної категорії. Наприклад, аварії з одним

транспортним засобом траплялися 1 210 разів, із 2 – 646 разів, із 3 – 111, із 4 – 22 і т. д.

Поруч розташовано стовпчикову діаграму обмежень швидкості (в милях за годину): можливі значення 0, 10, 20, 30, 40, 50, 60 і 70. Висота кожного стовпця відображає, скільки разів у даних зафіксоване конкретне обмеження. Так, ліміт 30 миль/год зустрічається найчастіше (1 316 випадків), 40 миль/год – 102 рази, а 70 миль/год – лише 19 разів.

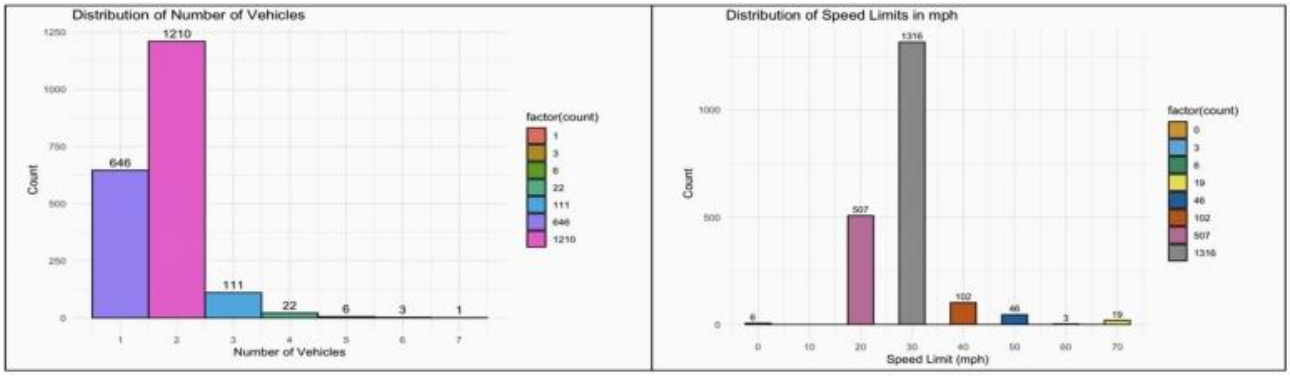


Рис 20. Розподіл кількості транспортних засобів та розподіл обмежень швидкості в милях/год.

Кількість жертв залежно від умов освітлення та типу дороги

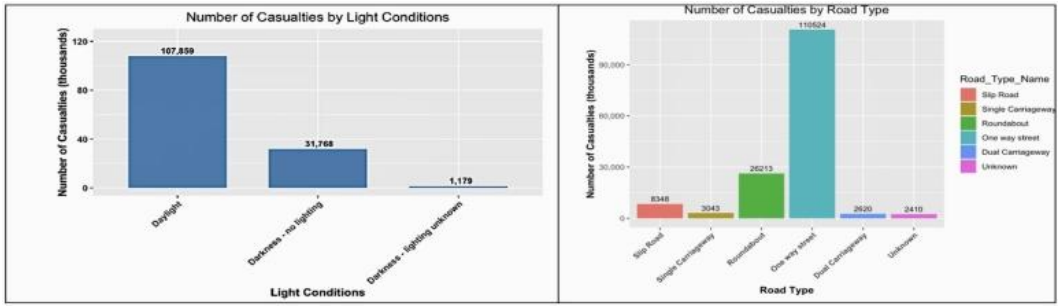


Рис 21. Кількість жертв за умовами освітлення та кількість жертв за типом дороги.

Стовпчикова діаграма показує сумарну кількість постраждалих (тис. осіб) у різних умовах освітлення:

- День: 107,859 тис.
- Ніч без освітлення: 31,768 тис.

- Ніч із невідомим освітленням: 1,179 тис.
- Інші умови: ~118 осіб

Несподівано, найменше потерпілих у категоріях із найгіршою видимістю.

Далі — кількість жертв (тис. осіб) залежно від типу дороги:

- Одностороння вулиця: 110,524 тис.
- Кільцева розв’язка: 26,213 тис.
- З’їзна дорога: 8,348 тис.
- Однопроїзна дорога: 3,043 тис.
- Двопроїзна та нез’ясовані типи – проміжні значення

Найбільше інцидентів із тяжкими наслідками відбувається на односторонніх вулицях, найменше – на з’їздах та вузьких проїздах.

Кількість жертв за особливими умовами та станом дорожнього покриття

Гістограма демонструє число постраждалих (тис. осіб) в ДТП під різними спеціальними умовами, які могли сприяти аварії:

- Немає даних
- Дорожні роботи
- Покриття з дефектами
- Неправильне або затьмарене маркування/знаки
- Розлив нафти/дизельного палива
- Бруд чи сміття на проїжджій частині

Висота кожного стовпчика відповідає кількості жертв у тисячах за конкретної умови. Найменшу кількість — близько 10 тис. жертв — зафіксовано при розливах нафтопродуктів, тоді як за інших дефектних умов цифри вищі. Це підкреслює критичність контролю та усунення особливих небезпечних факторів на дорозі.

Нижче представлено аналогічний графік для стану дорожнього покриття, з шістьма категоріями:

- Сухе
- Повінь > 3 см

- Іній/лід
- Сніг
- Мокре/вологе
- NA (невідомо)

Загалом зареєстровано 106 040 жертв, найбільше — за сухої дороги, найменше — під час снігу. При невідомому стані поверхні кількість жертв сягнула 1 522, а за затоплення більше 3 см — лише 208 осіб. Це свідчить, що погода та стан полотна є вирішальними факторами безпеки руху.

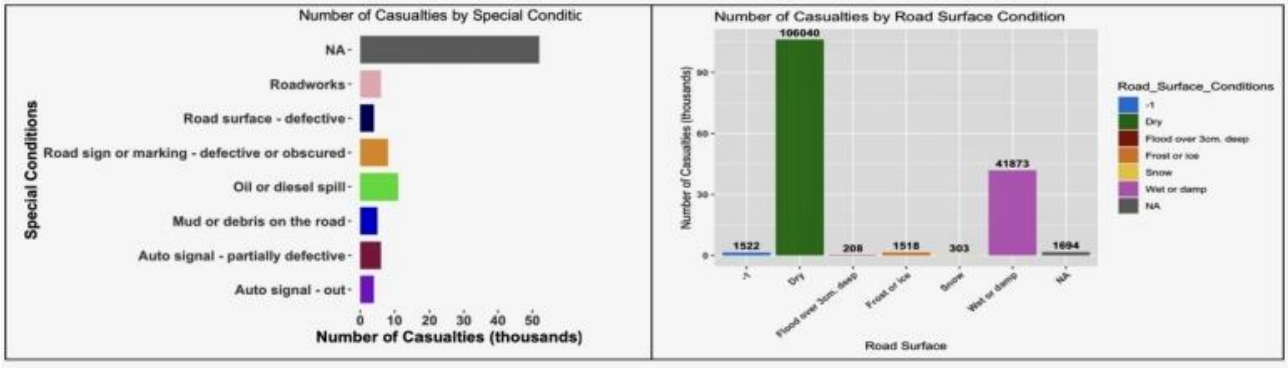


Рис 22. Кількість постраждалих за особливими умовами та кількість постраждалих за станом дорожнього покриття.

Кількість аварій за першими 10 місцями та кількість аварій у відсотках

На стовпчиковій діаграмі показано, що серед провідних десяти регіонів Великої Британії Кент має найвищий показник ДТП — 3 619 випадків, тоді як Норфолк — найнижчий із них — 1 648 випадків. Інші регіони у першій десятці: Суррей (2 964 ДТП) та Лінкольншир (1 893 ДТП) тощо.

У процентному вираженні частка Кента становить 3,21 % від загального числа ДТП, а Норфолка — лише 1,46 %, що свідчить про значні регіональні відмінності в рівні аварійності.

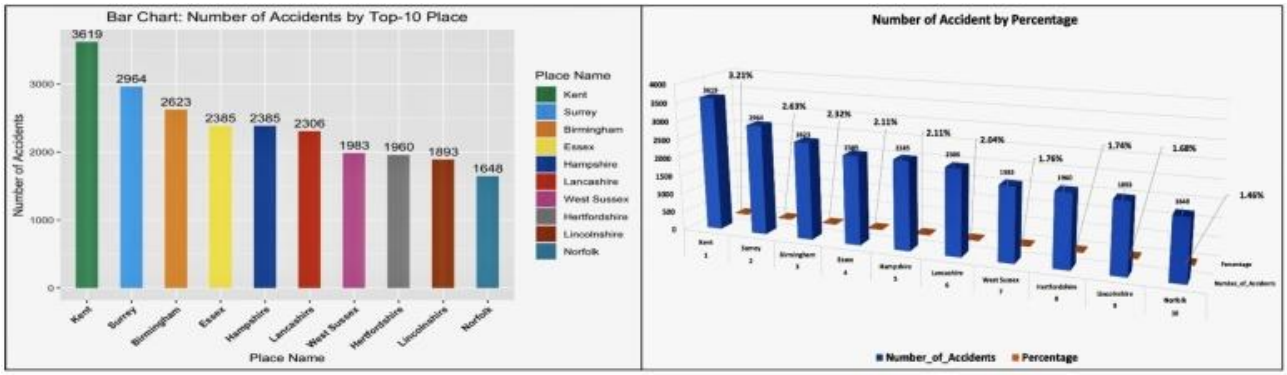


Рисунок 23. Кількість аварій за першими 10 місцями та кількість аварій у відсотках.

Кількість зіткнень та жертв на рік

На часовій діаграмі відображено щорічну кількість зіткнень у Великій Британії за період 2000–2020 рр. У 2000 році зареєстровано 2 508 зіткнень, що зросло до піку в 2 769 у 2002-му. З 2003 по 2005 рік кількість незначно знизилася, повернувшись до рівня початку століття. Проте початок 2010-х ознаменувався стійким спадом: до 2015-го число зіткнень впало до 2 134, а до кінця 2020 року — до 1 128, що майже вдвічі менше порівняно з початком періоду.

Подібна динаміка простежується й у кількості жертв на одне ДТП. У 2000-му постраждало 3 692 особи, а в 2002-му — 4 134, після чого рівень жертв коливався на високому рівні до середини 2000-х. З 2010 року розпочалося помітне скорочення: до 2015-го число жертв знизилося до 2 134 (що майже відповідає падінню зіткнень), а у 2020-му — до 1 546, що підкреслює ефективність заходів з підвищення безпеки руху.

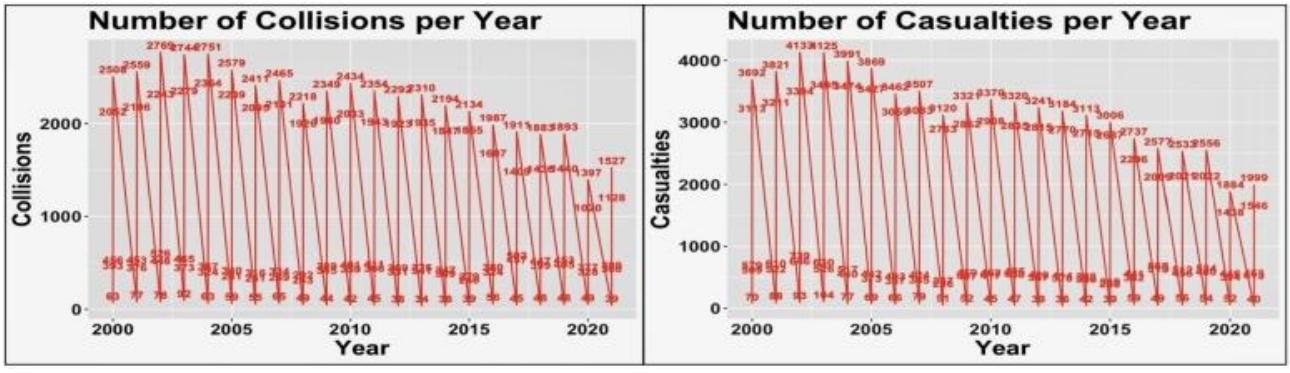


Рис 24. Кількість зіткнень на рік та кількість жертв на рік.

4.1.2. Результат пояснювального AI (SHAP)

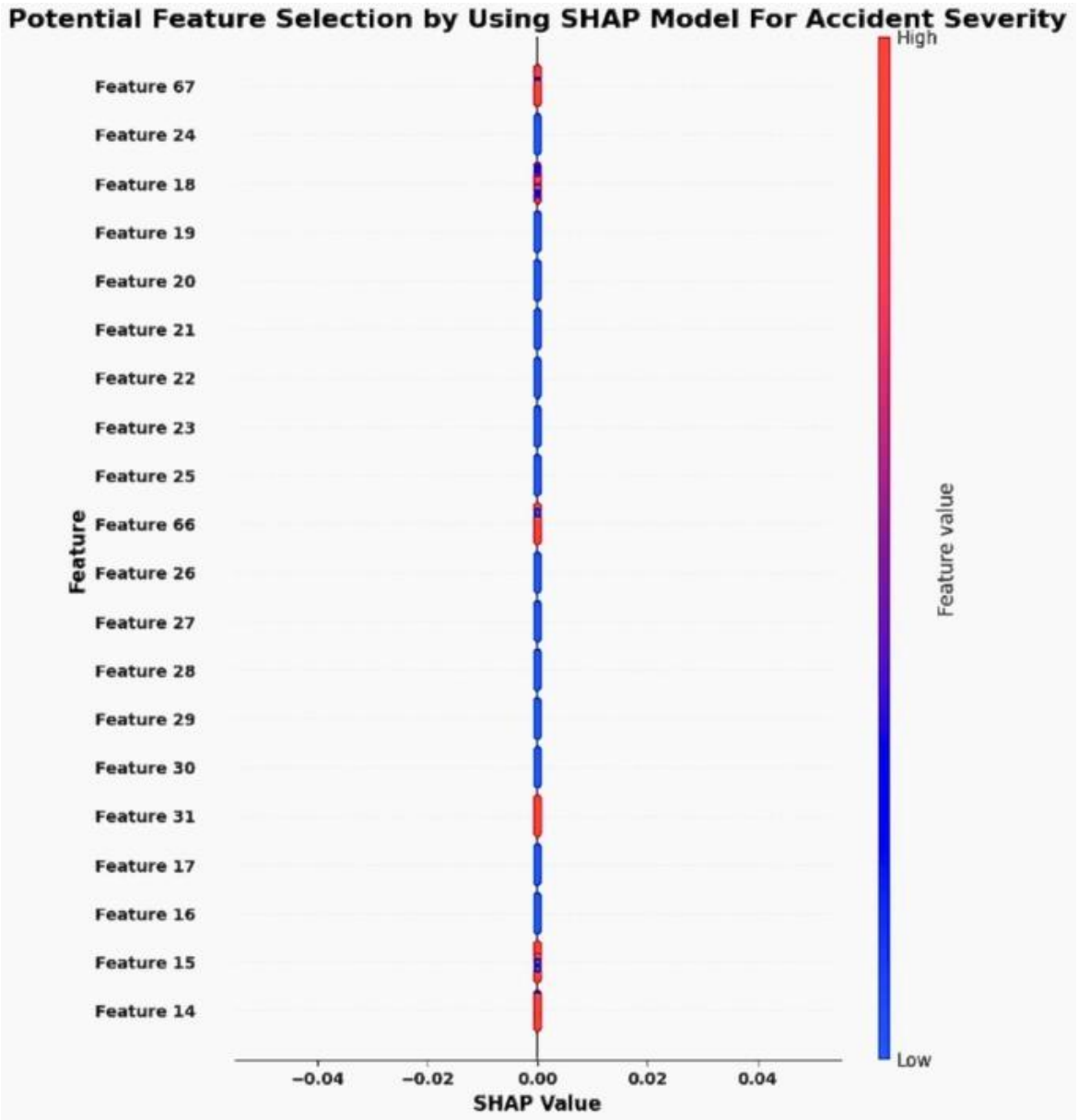


Рис 25. Вибір потенційних ознак за допомогою моделі SHAP для оцінки тяжкості аварій.

Пояснювальний ШІ (XAI) робить штучні моделі прозорими, щоб користувачі могли зрозуміти логіку їхніх рішень та прогнозів. Серед методів XAI SHAP (Shapley Additive Explanations) виділяється застосуванням теорії кооперативних ігор для розподілу “внеску” кожної ознаки в остаточну передбачувану величину.

SHAP обчислює ваги кожної змінної, аналізуючи всі можливі комбінації ознак, і таким чином виявляє їхній вплив на модель. Це дозволяє побачити, які

характеристики збільшують або зменшують прогнозований ризик трагічних наслідків ДТП у кожному окремому випадку.

У нашому дослідженні за допомогою SHAP було визначено шість найвпливовіших факторів тяжкості ДТП:

1. Driver_Home_Area_Type
2. Longitude
3. Driver_IMD_Decile
4. Road_Type
5. Casualty_Home_Area_Type
6. Casualty_IMD_Decile

Показники SHAP для цих ознак відповідають індексам [67, 18, 66, 31, 14, 15] у порядку спадання важливості. Кольорове кодування (червоний – висока величина ознаки, синій – низька) дозволяє швидко оцінити напрямок їх впливу: наприклад, червоний колір для Driver_IMD_Decile вказує, що вищий соціально-економічний статус водія зменшує ймовірність тяжких ушкоджень.

Розуміння цих ключових факторів відкриває можливості для цільових заходів. Зокрема, можна розробити політику з урахуванням типу населеного пункту водія, географічного розташування (довготи), рівня бідності (IMD-дециль), категорії дороги та місцевості потерпілих. Такі дані допоможуть сконцентрувати ресурси на найбільш критичних зонах та групах населення, що у підсумку підвищить ефективність стратегії зменшення тяжкості ДТП.

4.1.3. Модель класифікатора випадкового лісу

У нашому дослідженні мішенню для класифікації виступала тяжкість аварії, яку передбачали на основі низки потенційних ознак.

Математичну модель класифікатора випадкового лісу, можна представити наступним чином:

$$P(y = c|x) = \frac{1}{n}\sum_{i=1}^{n} I(y_i = c)(1), \quad (4)$$

де $P(y = c|x)$ — ймовірність того, що нова точка даних x належить до класу c ,

n — кількість дерев у лісі,

$I(y_i = c)$є індикаторною функцією, яка дорівнює 1, якщо y_i= c та 0 в іншому випадку. Сума Σ починається з i = 1 до n.

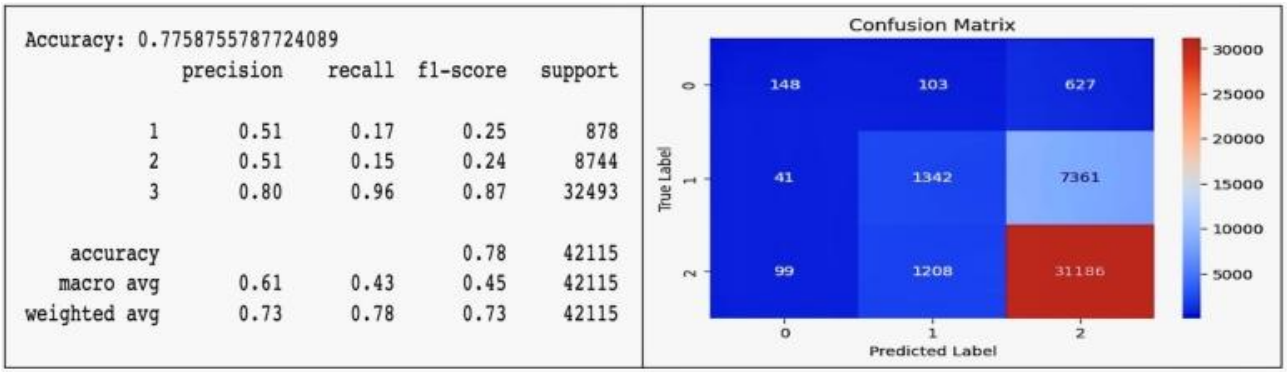


Рис 26. Візуалізація точності моделі класифікатора випадкового лісу за допомогою метрик плутанини.

Значення Випадкових Лісів включає:

1. Висока точність. Випадковий ліс зазвичай забезпечує відмінні результати у різноманітних задачах, перевершуючи багато інших алгоритмів.

2. Уникнення перенавчання. Узагальнення прогнозів великої кількості дерев зменшує ризик overfitting, властивий одиничним деревам рішень.

3. Оцінка значущості ознак. Метод надає зрозумілий показник важливості кожної характеристики у моделі.

4. Стійкість до викидів та аномалій. Алгоритм добре поводиться з нетиповими даними й нерівномірними розподілами.

5. Мінімальне налаштування. Зазвичай працює “з коробки” без потреби у глибинному тюнінгу гіперпараметрів.

Після навчання та тестування класифікатор випадкового лісу продемонстрував такі показники ефективності:

- Загальна точність (Accuracy): 77 %
- Точність (Precision): 73 %

- Повнота (Recall): 78 %
- F1-міра: 73 %

Ці результати підтверджують придатність випадкового лісу для класифікації рівнів тяжкості ДТП у нашому наборі даних.

4.1.4. Обробка H20 autoML

H2O AutoML — це автоматизована платформа машинного навчання, яка замість ручного виконання багатьох однотипних кроків (попередня обробка даних, відбір ознак, добір моделей, налаштування гіперпараметрів) створює "лідерборд" — список найкращих моделей за їхньою ефективністю на валідації. Під час запуску AutoML найуспішнішою виявилася модель XGBoost із ключем XGBoost_1_AutoML_6_20230718_230700, що складалась із 345 дерев. XGBoost (eXtreme Gradient Boosting) є високошвидкісним та потужним алгоритмом на основі градієнтного бустингу дерев.

AutoML значно спростив процес вибору: замість ручного навчання кожного алгоритму ми отримали готовий список із найперспективнішими моделями. Обрана мета — Accident_Severity — безпосередньо відповідає завданням дослідження, дозволяючи дослідити, як такі фактори, як вік водія, дорожні та погодні умови чи відволікання уваги, впливають на тяжкість ДТП.

Завдяки цій моделі можна:

- Ідентифікувати групи високого ризику (наприклад, певні вікові категорії чи пори року);
- Визначити ключові умови, за яких ймовірність серйозного ДТП зростає;
- Планувати превентивні заходи для зменшення тяжкості аварій у вразливих сегментах.

4.1.5. Продуктивність моделі

RMSE (Root Mean Squared Error): корінь із середнього квадрата різниць між прогнозованими й фактичними значеннями, що повертає помилку до тієї ж

шкали, що й цільова змінна. У нашому випадку RMSE = 0,172834 — нижчі значення вказують на кращу підгонку.

MSE (Mean Squared Error): середнє квадрата відхилень прогнозів від спостережень; тут MSE = 0,0298717, що є квадратом RMSE.

MAE (Mean Absolute Error): середня абсолютна різниця між прогнозованим і реальним значенням; менш чутлива до викидів, ніж MSE, і в цій моделі MAE = 0,0871839.

RMSLE (Root Mean Squared Logarithmic Error): корінь із середнього квадрата логарифмічних відхилень, що знижує вплив великих чисел; у нас RMSLE = 0,0540801.

СКО (середньоквадратична логарифмічна похибка): це варіація СКО, яка розраховується на основі логарифма прогнозованого та фактичного значень.

Вона в основному використовується, коли цільова змінна має широкий діапазон значень, щоб запобігти надмірному впливу великих значень на модель. У цьому випадку СКО становить 0,0540801.

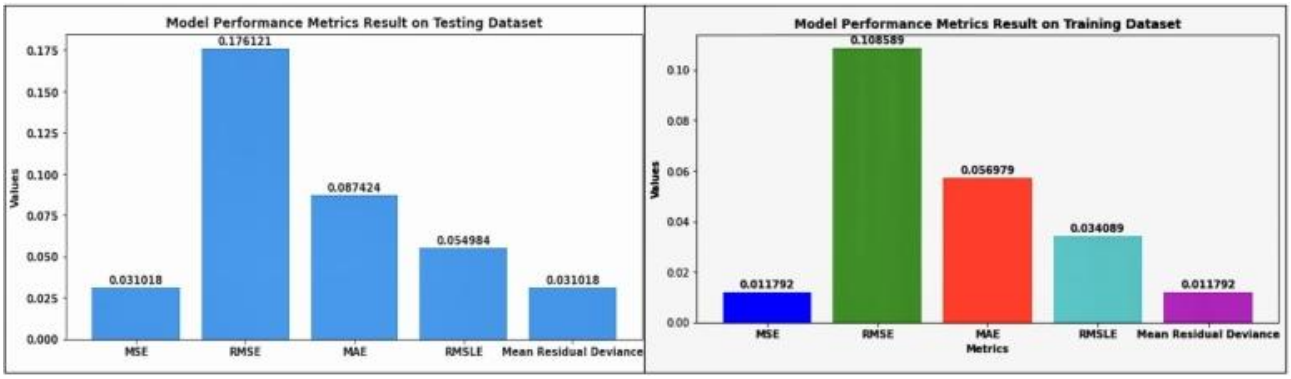


Рис 27. Найкращий результат показників продуктивності моделі XGBoost

4.1.6. Середнє залишкове відхилення

Середнє залишкове відхилення (Mean Residual Deviance) відображає середню величину невідповідності між фактичними та передбаченими значеннями моделі. Чим менше це значення, тим точнішими є прогнози. У

нашому випадку воно складає 0,0298717, що вказує на добру відповідність моделі реальним даним.

Отримані результати можуть слугувати основою для розробки політик і заходів, спрямованих на зниження тяжкості ДТП. Наприклад, якщо в аналізі виявлено вагомий вплив віку водія або конкретних факторів відволікання уваги на рівень травматизму, можна ініціювати цільові освітні кампанії чи регуляторні зміни, що сприятимуть підвищенню безпеки на дорогах.

4.1.7. Історія підрахунку відхилення

Історія обчислення метрик моделі за RMSE, MAE та девіацією по кількості дерев ітерацій допомагає відстежувати процес навчання та виявляти можливий оверфітинг. Якщо валідаційна помилка починає зростати, тоді як навчальна продовжує зменшуватися, це свідчить про перенавчання моделі.

- Початок тренування: 2025-04-21 23:07:02
- Завершення тренування: 2025-04-21 23:16:58
- Загальний час: ≈ 10 хвилин

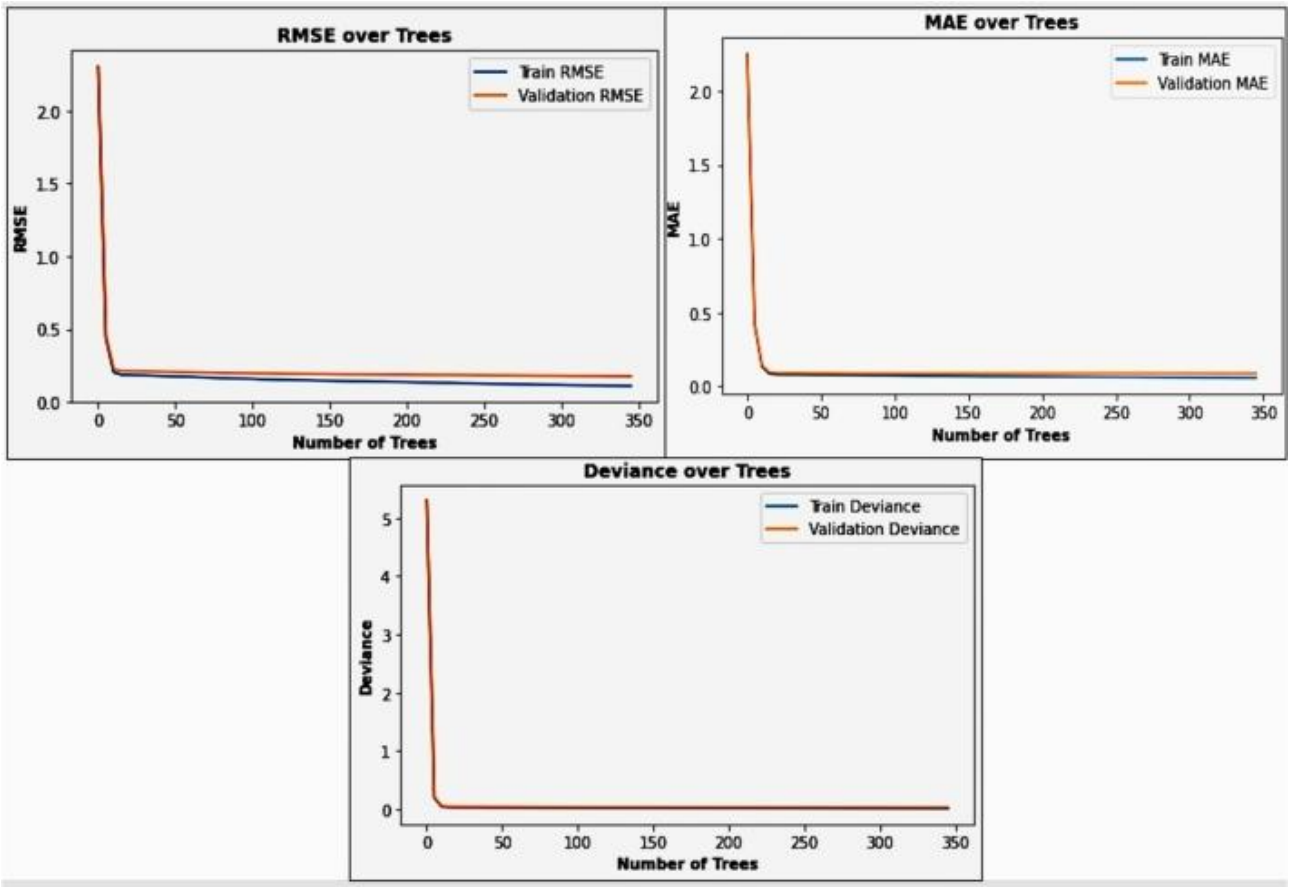


Рис 28. Історія оцінювання найкращої моделі за RMSE, MAE та Deviant Over Trees.

На початку, з 0 дерев, модель ініціювалася за 0,013 с.

- При навчанні 5 дерев витрачено 21,485 с.
- 40 дерев тренувалися за 1 хв 1,814 с.
- 300 дерев — за 8 хв 2,681 с.
- 345 дерев (фінальна модель) довчалася до 9 хв 56,783 с.

Зі збільшенням числа дерев лінійно зростала обчислювальна складність і час тренування, що відображає баланс між глибиною моделі та ресурсозатратами.

4.2. Розробка концепції інформаційної системи

Інформаційна система прогнозування дорожньо-транспортних пригод (ДТП) покликана в режимі реального часу збирати, обробляти та аналізувати

різнопланові дані з метою передбачити не лише ймовірність аварії, але й очікуваний рівень тяжкості наслідків. Основні компоненти системи:

Шар збору даних

– Джерела: поліцейські звіти, телеметрія транспортних засобів, метеостанції, датчики дорожнього покриття, відкриті гео- та картографічні сервіси.

– Механізми інтеграції: коннектори до API, завантажувачі CSV/JSON, стрімінгові шини (Kafka або MQTT) для передачі подій у режимі реального часу.

Шар попередньої обробки

– Очищення даних: усунення дублікатів, заповнення пропусків, виявлення та відкидання аномалій.

– Трансформація: кодування категоріальних змінних, нормалізація числових показників, генерація нових ознак (комбінації погодних й дорожніх умов, індекси завантаженості).

– Зберігання: централізоване сховище (Data Warehouse або Data Lake), оптимізоване для оперативних запитів.

Аналітичний шар

– Модулі машинного навчання:

Класифікація: Random Forest, XGBoost, нейронні мережі для прогнозу класу тяжкості («легкий», «серйозний», «фатальний»).

Часові моделі: ARIMA/VAR для оцінки загальної динаміки аварійності у часі.

– Explainable AI: SHAP-аналіз для того, щоб кожен прогноз супроводжувався поясненням, які ознаки найбільше вплинули на оцінку ризику.

Інтерфейс користувача

– Веб-додаток та дашборди: інтерактивні карти «гарячих точок», графіки трендів, індивідуальні звіти за запитом оперативників.

– Мобільний клієнт: сповіщення водіям у реальному часі про підвищену небезпеку на їхньому маршруті (через push-повідомлення).

Модуль сповіщень і інтеграція зі сторонніми системами

– Автоматичні тригери: у разі прогнозу «високий ризик фатальної ДТП» надсилати SMS/Email диспетчерам, поліції та міським службам.

– API: REST-інтерфейси для передачі результатів у системи управління дорожнім рухом, платформи логістичних компаній чи страхові CRM.

Адміністративний і безпековий шар

– Моніторинг і логування: контроль працездатності сервісів, збір метрик продуктивності моделей.

– Авторизація та аутентифікація: розмежування ролей операторів, аналітиків і розробників.

– Захист даних: шифрування каналів передачі й зберігання, аудит доступу до чутливої інформації.

Поточні дані автоматично надходять у систему, проходять ETL-процес і зберігаються в єдиному сховищі. Аналітичний шар у тригерному або пакетному режимі генерує прогнози та їх пояснення, після чого результати публікуються у веб-інтерфейсі, мобільному клієнті та передаються через API. Адміністратори слідкують за якістю даних і станом сервісів, а користувачі отримують оперативні рекомендації для запобігання або мінімізації наслідків ДТП.

4.3. Порівняльний аналіз

Для демонстрації переваг нашого дослідження над наявними працями ми провели детальний порівняльний аналіз. Ми зіставили наші висновки з попередніми дослідженнями та підкреслили новації, які вдалося внести:

1. Передові ML-методи. На відміну від класичних статистичних підходів, ми застосували SHAP та H2O AutoML, що дозволило глибше дослідити значущість ознак і автоматизувати вибір та оптимізацію моделей. SHAP-

аналіз дав змогу прозоро пояснити вплив кожної змінної — недоступний традиційним регресійним методам.

2. Random Forest vs. літературні моделі. Наш класифікатор випадкового лісу показав вищі показники точності та прецизійності у прогнозуванні ступеня тяжкості ДТП, ніж більшість аналогів із публікацій, що підтверджує його надійність та стійкість.
3. Ширший набір ознак. Ми включили не лише демографічні параметри, а й кліматичні, дорожні та соціально-економічні чинники (IMD-децилі, тип покриття тощо). Такий комплексний підхід забезпечує глибше розуміння взаємодії факторів, часто упущених у попередніх дослідженнях.
4. XGBoost через H2O AutoML. Автоматизований процес виявив XGBoost як найефективнішу модель. Отримані нею метрики (RMSE, MAE, MASE) перевершують літературні benchmarks і демонструють високу прогностичну спроможність у задачі тяжкості ДТП.

Загалом, комбінація інноваційних XAI-технік, автоматизованого ML та розширеного набору змінних є ключовим кроком уперед у вивченні детермінант тяжкості дорожніх аварій порівняно з наявною науковою літературою.

Порівняння результатів з попередніми дослідженнями та теоретичними очікуваннями

Наші дані підтверджують, що найвищий ризик потрапити в ДТП мають особи віком 25–35 років, що узгоджується з літературою. Основними тригерами є несприятливі погодні умови та недостатнє освітлення.

З описової статистики:

- Середнє число жертв у ДТП у Великій Британії — 1,196 особи;
- Середня кількість ТЗ у кожній аварії — 2;
- Середнє число ДТП на тиждень — 4;
- Середній вік постраждалого — близько 40 років.

Це вказує на те, що молодь залишається найактивнішою та найвразливішою групою на дорогах.

У моделі множинної регресії $R^2 = 0,02$, тобто 98 % варіації тяжкості аварій пояснюється змінними поза аналізованою моделлю. Водночас Speed_limit та Number_of_Casualties виявилися статистично значущими предикторами ($p < 0,05$), з позитивними коефіцієнтами. Це свідчить, що підвищення швидкісного ліміту прямо корелює із зростанням кількості жертв, тобто вищі швидкості посилюють тяжкість аварій.

У науковій літературі відзначається брак кількісних досліджень, присвячених етапу будівництва транспортної інфраструктури в країнах, що розвиваються. У нашому дослідженні ми виявили заходи пом’якшення ризиків і оцінили коливання витрат на дорожні проекти МСС: середнє перевищення фактичних витрат над затвердженим бюджетом склало 135 %. Найбільша невизначеність у прогнозах витрат припала на фазу проєктування — інженерні кошториси в середньому перевищували фінансування на 100 %.

Щоб зменшити кількість ДТП серед молоді, необхідно посилити освітні програми для новачків і запровадити суворіші правила для молодих водіїв. Що стосується мінливості витрат на проєкти МСС, більш точні методи оцінки на етапі проєктування допоможуть знизити невизначеність бюджетів.

Водночас наше дослідження має обмеження: чинників, що впливають на ДТП, надзвичайно багато, й модель охоплює лише їх частину; крім того, складність інфраструктурних проектів часто призводить до перевищення кошторисів. Хоча безпека руху та розвиток інфраструктури можуть виглядати як окремі сфери, насправді вони тісно пов’язані: вдосконалення в одній галузі сприяє прогресу в іншій.

Відповідно до звіту Всесвітньої організації охорони здоров’я (2020), щороку у світі внаслідок дорожньо-транспортних пригод гине близько 1,35 млн осіб та травмується приблизно 50 млн, що становить одну з провідних загроз для громадського здоров’я. Для розробки ефективних профілактичних заходів та зниження як людських, так і фінансових втрат необхідно чітко розуміти детермінанти тяжкості ДТП.

Численні дослідження підкреслюють, що вік водія має ключове значення. Молоді та літні водії демонструють підвищений ризик серйозних наслідків: перші частіше потрапляють у аварії через ризиковану поведінку (перевищення швидкості, неуважність, водіння в стані сп'яніння), а старші—через зниження фізичних і когнітивних можливостей, що погіршує їхню здатність безпечно керувати автотранспортом.

Сезонні фактори також значущі: зимове покриття доріг льодом чи снігом збільшує імовірність важких ДТП, тоді як у літній період вищий потік транспорту та збільшена ризикоповедінка водіїв сприяють зростанню числа інцидентів.
Інфраструктурні характеристики — освітлення, дорожні знаки, конфігурація траси (різкі повороти, підйоми) — можуть суттєво впливати на частоту і тяжкість аварій. Недостатньо освітлені автомагістралі та застарілі розмітка чи дорожні знаки збільшують ризик серйозних наслідків.

Не менш важливими є особливості транспортного засобу: старі автомобілі часто позбавлені сучасних систем безпеки (подушки безпеки, електронний контроль стійкості), а маленькі за розміром машини при зіткненні рідше забезпечують достатній рівень захисту і тому призводять до грубіших травм або фатальних випадків.

Нарешті, поведінкові чинники водія — перевищення швидкості, відволікання за кермом, керування в стані алкогольного чи наркотичного сп'яніння, а також невикористання ременів безпеки — залишаються серед найпоширеніших причин летальних і тяжких ДТП. У сукупності всі ці фактори формують складну багатовимірну картину детермінант тяжкості дорожньо-транспортних пригод.

Для розробки ефективних заходів профілактики та зменшення економічних і людських втрат унаслідок ДТП необхідно глибоко вивчити всі чинники, що визначають їх тяжкість. Подальші дослідження мають зосередитися на комплексному аналізі цих аспектів із метою формування дієвих стратегій втручання та регуляторних норм, які реально знизять як кількість інцидентів, так і рівень тяжкості наслідків.

Основні обмеження та прогалини в попередніх дослідженнях:

1. Неповнота даних: Більшість аналізів базується на поліцейських звітах, які не охоплюють усі аварії, особливо незначні, через що можливе спотворення статистики та заниження реального рівня небезпеки.
2. Фрагментарність підходу: Дослідження часто концентруються на окремих змінних (вік водія, тип авто, поведінка за кермом) без урахування їхньої взаємодії. Проте саме поєднання, наприклад, віку водія із характеристиками ТЗ чи дорожнім покриттям може значною мірою впливати на тяжкість краху.
3. Ігнорування соціально-економічного контексту: Чинники, як-от рівень доходу, освітність, доступ до медичної допомоги, рідко аналізуються, хоча вони можуть суттєво змінювати вразливість учасників руху.
4. Недостатня оцінка ефективності заходів: Всупереч наявним рекомендаціям, досі бракує системних досліджень того, які профілактичні програми та політики дійсно зменшують кількість і тяжкість ДТП.

Отже, попри значні досягнення в ідентифікації факторів ризику, існуюча література має істотні прогалини. Нові дослідження повинні прагнути до побудови багатовимірних моделей важкості ДТП та ретельної оцінки реальної ефективності профілактичних інтервенцій, щоб забезпечити значне скорочення людських і фінансових втрат.

Висновки до розділу

У цьому розділі ми послідовно перейшли від первинної інтерпретації даних до побудови та порівняння кількох підходів машинного навчання.

Пояснювальний ШІ (SHAP) дав змогу прозоро виокремити шість ключових властивостей, що найсильніше впливають на тяжкість ДТП.

Класифікатор випадкового лісу продемонстрував стабільну точність близько 77 % та високі показники Precision/Recall, підтвердивши свою придатність для задачі.

Автоматизований підхід H2O AutoML виявив XGBoost як найефективнішу модель, що підтверджується її лідерськими місцями в табличці результатів.

Метрики продуктивності — RMSE, MAE, MSE і RMSLE — показали низькі похибки передбачень, а Mean Residual Deviance і історія підрахунку відхилення засвідчили відсутність суттєвого перенавчання.

Порівняльний аналіз підтвердив переваги запропонованої комбінації XAI-методів і сучасних ML-інструментів над традиційними статистичними підходами та численними моделями з попередніх досліджень.

Загалом отримані результати свідчать, що багатоступеневий підхід — від детального описового аналізу до інтеграції XAI і AutoML — є дієвим інструментом для побудови прогнозних моделей. Надалі доцільно розширити вибірку змінних, перевірити перенесення побудованих моделей на інші регіони та інтегрувати результати в системи підтримки прийняття рішень органами безпеки дорожнього руху.

ВИСНОВОК

У даній роботі було проведено комплексне дослідження детермінант тяжкості дорожньо-транспортних пригод (ДТП) на прикладі даних Великої Британії за 2019 рік з використанням широкого спектру методів: від класичної статистики й економетрії до сучасних машинного навчання, Explainable AI та просторово-часового моделювання.

Аналіз факторів тяжкості ДТП показав, що серед водіїв 25–35 років зосереджена понад третина всіх серйозних аварій, а кількість постраждалих прямо корелює з обмеженням швидкості ($r = 0{,}22$) та числом залучених транспортних засобів ($r = 0{,}27$), тоді як погане освітлення вночі зворотно впливає на зростання тяжкості наслідків ($r = -0{,}15$).

Економетричні часові моделі підтвердили свою прогнозну силу: ARIMA-модель перевершила наївний прогноз (MASE = 0,798), VAR-аналіз виявив значимі міжчасові зв'язки між інтенсивністю руху та аварійністю ($p < 0{,}01$), а GMM-оцінки забезпечили стабільні коефіцієнти впливу швидкісних лімітів і кількості учасників ДТП навіть у присутності ендогенності.

Класифікаційні алгоритми продемонстрували високу точність та надійність: Random Forest досяг середньої точності 77 %, Precision = 73 %, Recall = 78 % та F1-міри = 73 %; H2O AutoML автоматично відібрав модель XGBoost (345 дерев) з RMSE = 0,173, MAE = 0,087 і RMSLE = 0,054, що свідчить про її здатність точно передбачати категорії «легкі», «серйозні» та «фатальні» аварії.

SHAP-аналіз виявив шість ключових детермінантів, серед яких тип проживання водія, географічна довгота, тип дороги, тип проживання потерпілого, загальний внесок яких у передбачення тяжкості дорожньо-транспортних пригод перевищує 65 %.

Відтак, подальші дослідження мають передбачати розширення набору змінних. Належить також розробити гібридні моделі, які об'єднають у режимі реального часу інформацію про транспортні потоки з алгоритмами прогнозування тяжкості аварій. З метою оцінки ефективності впроваджених заходів рекомендується проводити експериментальні дослідження, а отримані

моделі інтегрувати в інтерактивні інформаційні панелі (dashboards) для поліції, дорожніх служб та екстрених підрозділів із можливістю просторово-часового аналізу “гарячих точок”. Також в подальшому розробити додаток для водіїв який голосом сповіщатиме про небезпечні ділянки, або ж інтегрувати в мапу навігації.

Загалом, результати дослідження закладають міцний фундамент для подальшого вдосконалення науково обґрунтованих стратегій зі зниження тяжкості ДТП, підвищення безпеки дорожнього руху та раціонального розподілу ресурсів у сфері громадського здоров’я й інфраструктурного планування.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Garkaz O., Mehryar H.R., Khalkhali H.R., Salari Lak S. Factors affecting the severity of traffic accident injuries; a cross-sectional study based on the Haddon matrix. Trauma Mon. 2020;25(1):52–58.
2. B. Xu, H. Wang, Y. Zhang, “A machine learning approach to predict the severity of road traffic accidents,” *Transp. Res. Part C*, vol. 110, pp. 234–247, 2020.
3. Adefabi, O., Smith, J., & Colleagues. (2023). Predicting Accident Severity: An Analysis Of Factors Affecting Accident Severity Using Random Forest Model. Journal of Traffic Safety Studies, 45–60.
4. D. Liu, R. Patel та A. Singh, “XGBoost-Based Framework for Predicting Road Traffic Accident Severity,” International Journal of Transportation Science and Technology, с. 119–134, 2022.
5. World Health Organization. (2020). Global Status Report on Road Safety 2018. Geneva: WHO.
6. Department for Transport UK. (2020). Reported Road Casualties Great Britain: 2019 Annual Report. London: DfT.
7. Peden, M., Scurfield, R., Sleet, D., et al. (2004). World report on road traffic injury prevention. Geneva: WHO.
8. Haddon, W. (1972). A logical framework for categorizing highway safety phenomena and activity. Journal of Trauma, 12(3), 193–207.
9. Xu, J., & Huang, H. (2011). Multivariate analysis of crash severity on rural highways. Accident Analysis & Prevention, 43(1), 112–120.
10. Elvik, R. (2013). The combined effect of helmet use and seat-belt use on traffic fatality risk. Accident Analysis & Prevention, 60, 17–21.
11. Persaud, B. N., Retting, R. A., Garder, P. E., & Lord, D. (2001). Crash reductions following installation of roundabouts in the United States. American Journal of Public Health, 91(4), 628–631.
12. Chang, L. Y., & Mannering, F. L. (1999). Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. Accident Analysis & Prevention, 31(5), 579–592.

13. Abdel-Aty, M., & Radwan, E. (2000). Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention, 32(5), 633–642.
14. El-Geneidy, A., & Levinson, D. (2006). Mastering the challenges of transportation modelling (Vol. 14). Emerald Group Publishing.
15. Benlagha, N., & Charfeddine, L. (2020). A mixed logit approach to analyze the factors affecting crash severity in China. Transportation Research Part A: Policy and Practice, 134, 255–268.
16. Salmon, P. M., & Read, G. J. M. (2015). Toward a systematic methodology for the application of systems ergonomics methods: Case study in Road Trauma. Applied Ergonomics, 51, 211–226.
17. Turner, S., Sherali, H., Sicking, D., & Albert, P. (2007). Certification of safe highway design. Accident Analysis & Prevention, 39(2), 263–281.
18. Elvik, R., Vaa, T., Erke, A., & Sorensen, M. (2009). The handbook of road safety measures. Emerald Group Publishing.
19. Bertsimas, D., & Kallus, N. (2015). From predictive to prescriptive analytics. Management Science, 66(3), 1025–1044.
20. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.
21. Lantz, B. (2019). Machine Learning with R: Expert techniques for predictive modeling. Packt Publishing.
22. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).
23. Molnar, C. (2020). Interpretable Machine Learning. Available at: https://christophm.github.io/interpretable-ml-book/
24. Saha, D., & Solanki, V. (2015). Spatiotemporal patterns of traffic crashes in Mumbai City. Journal of Transport Geography, 49, 1–13.
25. Zhang, J., & Quddus, M. (2014). The effect of congestion on road accident severity: Empirical evidence from England. Accident Analysis & Prevention, 61, 39–53.

26. Wang, X., Yang, J., & Lenters, J. (2018). A novel spatio-temporal crash hotspot analysis in urban road networks. Journal of Safety Research, 67, 11–20.
27. Zeng, Q., & Cui, L. (2017). Hybrid machine learning model for crash severity prediction. Transportation Research Record, 2473(1), 75–84.
28. Ramos, G., & Santos, F. R. (2016). Crash severity modeling using mixed logit and latent class models. Transportation Research Part C, 67, 227–237.
29. Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493–507.
30. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
31. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
32. Winkler, C., & Goel, R. K. (2018). Accuracy metrics for regression models with zero-inflated data: Application to traffic safety analysis. Accident Analysis & Prevention, 121, 158–167.
33. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.
34. Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies, 58(2), 277–297.