НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Кваліфікаційна наукова
праця на правах рукопису

ПАНІБРАТОВ РОМАН СЕРГІЙОВИЧ

УДК 004.89+519.22]:368.025.6(043.3)

ДИСЕРТАЦІЯ

СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ АНАЛІЗУ АКТУАРНИХ РИЗИКІВ

122 – Комп’ютерні науки
12 – Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії.

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело _____ Р.С. Панібратов

Науковий керівник: Бідюк Петро Іванович, доктор технічних наук, професор

Київ – 2026

АНОТАЦІЯ

Панібратов Р. С. Система підтримки прийняття рішень для аналізу актуарних ризиків. — Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 “Комп’ютерні науки” — Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, 2026.

Термін “актуарний ризик” визначається як тип ризику, коли висунуті гіпотези актуаріями щодо ціноутворення страхових полісів можуть бути неточними або некоректними. Рівень даного типу ризику прямо залежить від ступеню стійкості припущень, утворених внаслідок моделей задавання цін, які використовуються страховими організаціями під час встановлення премій. Безумовно, поняття страхування є впливовим фактором ризику та бізнесу всіх страховиків. Оскільки оформлення страхових полісів включає в себе величезний набір аспектів, то сама задача стає по собі доволі складною. Актуарії зазвичай є фахівцями зі справами, що стосуються виявлення ризиків, отримання цінної інформації та розуміння складності ситуацій. Вони є тими, на яких страхові компанії покладають великі надії для аналізу та керування ризиками.

Враховуючи динамічність ризиків, вони, очевидно, створюють проблеми під час актуарного моделювання. Щоб актуарії могли коректно фіксувати та оцінювати ризики, їм постійно потрібно модифікувати існуючі методи або розробляти абсолютно нові підходи. Задача моделювання має бути розв’язана з метою демонстрації та пропозиції коректних оцінок ризиків. Часто традиційні актуарні підходи не є ефективними через те, що ризики не містять важливих історичних даних. Крім того моделювання сукупного ефекту ризиків також є непростим завданням, оскільки групи ризиків можуть мати складні зв’язки або бути взаємозалежними. Через постійну нестачу або недоступність даних та динамічність факторів ризику традиційні показники ризику зазвичай містять

неточні результати. Для аргументування припущень та оцінок актуаріям зазвичай доводиться застосовувати аналіз чутливості і методи протидіям невизначеності. Основною метою є створення спеціалізованої системи підтримки прийняття рішень, що дає можливість більш детально аналізувати актуарні ризики, давати оцінку можливим майбутнім втратам та їх ймовірність. Практична цінність роботи полягає в створенні інтелектуальної системи підтримки прийняття рішень для аналізу актуарних ризиків. Результати роботи доведено до інженерного рівня і впроваджено в Інституті телекомунікацій і глобального інформаційного простору Національної академії наук України з метою автоматизованого розв'язання задач моделювання, аналізу та прогнозувань втрат страхових компаній, як випадковий процес. Спроектовано програмний продукт для оцінки можливих втрат в грошових одиницях та їх ймовірність. Всі теоретичні та практичні результати дисертаційної роботи у повній мірі опубліковано у фахових наукових виданнях, що входять до відповідного встановленого переліку, а також виконано їх належну апробацію на наукових конференціях і семінарах.

У дисертаційній роботі проаналізовані властивості страхування, поняття страхового ринку і його учасників. Проведено аналіз поточного стану страхового ринку України та проблеми, з якими він зіштовхується на даний момент. Виявлено нерозривний зв'язок між страховою діяльністю та страховим ризиком. Водночас було відмічено, що при дослідженні ризиків у страхуванні дуже мало уваги приділяється якісним методам оцінювання. Звідси виникає потреба у комбінації економічних та математичних алгоритмів моделювання та аналізу фінансових ризиків для розрахунку оптимального значення страхових премій за знайденою ймовірністю банкрутства. Проведений аналіз існуючих методів оцінювання ризиків у страхуванні показав, що їх усі можна поділити на дві групи: високорівневі моделі, що ґрунтуються на аналізі результатів та низькорівневі моделі, що ґрунтуються на дослідженні чинників ризику. Також підходи оцінок ризиків можна класифікувати на методи в умовах повної визначеності, часткової та повної невизначеності. Виконано аналіз по

відношенню до типів даних та видів невизначеності. Крім того були виділена група невизначеностей, що розглядалася у подальших експериментах. Розглянуто основні числові характеристики, що використовуються при аналізі вибірок даних. Проведено аналіз існуючих рішень технологій аналізу актуарних ризиків, що використовуються станом на зараз.

Після проведення аналізу було прийнято рішення використовувати Байєсівську методологію та узагальнені лінійні моделі при розробці системи підтримки прийняття рішень. Мережі Байєса характеризуються тим, що вони можуть використовувати як статистичні дані так і експертні оцінки і відображають процес прийняття рішень в режимі реального часу. Також даний підхід дозволяє враховувати різноманітні види невизначеностей. Для обчислювання ймовірнісного висновку був застосований точний метод Lauritzen-Spigielharter або LS-метод, що ґрунтується на методах кластеризації. Крім того були розглянуті приклади застосування мереж Байєса у сфері страхування. Вибір узагальнених лінійних моделей ґрунтується тим, що вони дозволяють оминати обмеження класичних моделей лінійної регресії та дозволяє обирати найважливіші змінні. Розглянуто основні складові даного підходу і основні методи аналізу залишків та діагностики моделей. Крім того були розглянуті і альтернативні підходи як методи нечіткої логіки: алгоритм Мамдані і алгоритм Сугено.

Також було відмічено, що при аналізі ризиків також потрібно враховувати і різні типи невизначеностей. Їх неможливо повністю усунути, але можна мінімізувати до певного рівня. Невизначеності набагато легше виявляти під час попередньої оброки даних, тому було проведено огляд існуючих методів як числових так і категорійних даних. Крім того були розглянуті методи вибору підмножини важливих ознак, як рекурсивне виключення ознак, тест Хі-квадрат і метод випадкового лісу. Також було розглянуто проблеми відсутності та зашумленості даних. Важливим етапом побудови узагальнених лінійних моделей є оцінювання параметрів. Для цього було використано рекурсивний ітераційно-зважувальний рекурентний метод найменших

квадратів, метод адаптивної оцінки моментів або Adam і метод Монте-Карло для марківських ланцюгів. Проведено огляд щодо ітеративних і неітеративних реалізацій останнього підходу. При аналізі невизначеностей в актуарних процесах було запропоновано використовувати фільтр Калмана для мінімізації статистичної невизначеності. Оскільки страхові дані не завжди є в публічному доступі, то було використано підхід імітаційного моделювання для генерації даних. При оцінюванні параметрів узагальнених лінійних моделей разом з попередньою обробкою даних було розглянуто випадки застосування та без застосування фільтра Калмана. Результати дослідження продемонстрували, що метод Adam та Монте-Карло продемонстрували непогані результати. Останній алгоритм використовувався у подальшому, оскільки він здатний до адаптації у разі надходження нових даних.

Розроблено інтелектуальну систему підтримки прийняття рішень, що дозволяє аналізувати актуарні ризики, а саме ризики втрат для страхової компанії, за допомогою узагальнених лінійних моделей і мереж Байєса. В якості даних для програми використовувалися страхові дані клієнтів Singapore Actuarial Society. Також при розробці програмного продукту продемонстровано поетапну реалізацію LS-методу і порівняно результати ймовірнісних висновків з результатами логістичної регресії. Дослідження показали, що мережа Байєса у більшості випадків продемонструвала набагато кращі результати, ніж класичний метод класифікації. Створена система підтримки прийняття рішень показала непогані результати, що свідчить про те, що програма здатна до детального аналізу ризику втрат.

Наукова новизна одержаних результатів полягає у такому:

— вперше побудовано та досліджено вибрані типи узагальнених лінійних моделей для аналізу актуарних процесів, які відрізняються можливістю їх застосування для формального опису лінійних і нестаціонарних нелінійних процесів і забезпечують можливість оцінювання актуарного ризику з високою якістю;

— вперше побудовано нову модель на основі LS-методу у формі байєсівської мережі для оцінювання актуарних ризиків, яка характеризується коректністю формального опису вибраного процесу за байєсівським інформаційним критерієм і забезпечує оцінювання високоякісних прогнозів актуарних процесів;

— вперше створено і реалізовано програмно системну методологію побудови адаптивних моделей актуарних процесів, яка відрізняється запропонованим методом моделювання в умовах наявності статистичних невизначеностей даних, методом оцінювання параметрів моделей і забезпечує підвищення адекватності побудованих моделей та якості оцінок прогнозів;

— удосконалено метод оцінювання параметрів узагальнених лінійних моделей завдяки комплексному застосуванню методу Монте-Карло для марківських ланцюгів, методу Adam та ітераційного рекурсивно-зважувального методу найменших квадратів, що гарантує підвищення якості параметрів.

Ключові слова: нелінійні процеси, мережі Байєса, LS-метод, узагальнені лінійні моделі, логістична регресія, рекурсивний ітеративно-зважувальний рекурентний метод найменших квадратів, метод Adam, метод Монте-Карло для ланцюгів Маркова, фільтр Калмана, імітаційне моделювання, система підтримки прийняття рішень.

ABSTRACT

Panibratov R.S. Decision support system for analysis of actuarial risks. — Qualifying scientific work on the rights of the manuscript.

Dissertation for obtaining a scientific degree of a Ph. D. degree in specialty 122 “Computer Science”. — National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ministry of Education and Science of Ukraine, Kyiv, 2026.

The term “actuarial risk” is defined as a type of risk, when hypotheses, suggested by actuaries regarding the pricing of insurance policies can be inaccurate or incorrect. The level of such type of risk directly depends on stability of the assumptions, formed as a result of the pricing models, used by insurance companies when adjusting premiums. Of course the concept of insurance is an influential factor in the risk and all types of businesses of insurers. Since the design of insurance policies includes a large set of aspects, the task becomes very complex itself. Actuaries are usually specialists in cases, which are related to identification of risks, obtaining valuable information and understanding the complexity of situations. They are the ones whom insurance companies place great hopes about risks analysis and management.

Considering the dynamism of risks, they obviously pose problems during actuarial modeling. In order for actuaries to correctly detect and evaluate risks, they constantly need to modify existing methods or develop completely new approaches. The modeling task must be solved in order to demonstrate and suggest correct risk assessments. Often, traditional actuarial approaches are note effective because risks do not contain important historical data. In addition, modeling of cumulative effect of risks is also a difficult task, since group of risks can have complex relationships or be interdependent. Due to the constant lack or unavailability of data and dynamism of risk factors, traditional risk indicators usually contain inaccurate results. To justify assumptions and estimates, actuaries usually have to apply sensitivity analysis and method to counteract uncertainties. The main goal is to create a specialized decision

support system that allows for a more detailed analysis of actuarial risks, to assess possible future losses and their probabilities. The practical significance of the work lies in the creation of an intelligent decision support system for analysis of actuarial risk. The results of the thesis fulfillment were brought to an engineering level and were used in the Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine in order to automatically solve the problems of modeling, analysis and forecasting of losses of insurance companies, as a stochastic process. A software products was designed to assess possible losses in monetary units and their probability. All theoretical and practical results of the dissertation published in professional domestic journals included in the relevant list, as well as their proper approbation at scientific conferences and seminars.

The dissertation analyzes the properties of insurance, the concept of insurance market and its participants. The current state of the insurance market of Ukraine and the problems it faces at the moment are analyzed. An extricable link between insurance activity and insurance risk is revealed. At the same time, it was noted that during studying risks in insurance, very little attention is paid to qualitative assessment methods. Hence the need for a combination of economic and mathematical algorithms for modeling and analyzing financial risks to calculate the optimal value of insurance premiums based on the found probability of bankruptcy. The analysis of existing methods for assessing risks in insurance demonstrated the that they all can be divided into two groups: high-level models based on the analysis of results and low-level models based on the study of risk factors. Also risk assessment approaches can be classified according to methods under conditions of complete certainty, partial and full uncertainty. An analysis was performed in relation to data types and types of uncertainty. In addition, a group of uncertainties was identified, which was considered in further experiments. The main numerical characteristics used in the analysis of data samples were considered. An analysis of existing solutions of actuarial risk analysis technologies used at the present time was conducted.

After the analysis, it was decided to use Bayesian methodology and generalized

linear models in the development of a decision support system. Bayesian networks are characterized by the fact that they can use both statistical data and expert estimations and reflect the decision-making process in real time. This approach also allows to take into account various types of uncertainties. To calculate the probabilistic inference, the exact Lauritzen-Spigielharter method or LS method, based in clustering methods, was applied. In addition, examples of application of Bayesian networks in the insurance sector were considered. The choice of generalized linear models is based on the fact, they they allow to bypass the limitation of the classical linear regression models and allow to choose the most important variables. The main components of this approach and the main methods of residual analysis and model diagnostics were considered. In addition alternative approaches were considered as fuzzy logic methods: the Mamdani algorithm and the Sugeno algorithm.

It was also noted that during the risk analysis, it is also necessary to consider various types of uncertainties. They cannot be completely eliminated, but they can be minimized to a certain level. Uncertainties are much easier to detect during preliminary data collection, so a review of existing methods for both numerical and categorical data was conducted. In addition, methods for selecting a subset of important features were considered, such as recurrent feature elimination, the Chi-square test and random forest method. The problems of missing and noisy data was also considered. An important stage in building generalized linear models is parameter estimation. For this task, the iteratively-reweighted least squares method, the adaptive moment estimation method or Adam, and the Monte-Carlo method for Markov chains were used. A review was conducted of iterative and non-iterative implementation of latter approach. During the analysis of uncertainties in actuarial processes, it was suggested to use the Kalman filter to minimize statistical uncertainty. Since insurance data is not always publicly available, a simulation modeling approach was applied to generate data. When estimating the parameters of generalized linear models, together with data preprocessing methods, the use of and without the Kalman filter was considered. The results of the experiment demonstrated that Adam method and Monte-Carlo methods showed acceptable results. The latter

algorithm was used in further steps, because it is capable of adaptation in the case of the new data.

An intellectual decision support system was developed, that allows analyzing actuarial risks, namely the risk of losses for an insurance company, by using generalized linear models and Bayesian networks. Insurance data of clients of the Singapore Actuarial Society were used as dataset for the program. Also, when developing the software product, step-by-step implementation of the LS-method was demonstrated and result of probabilistic inference were compared with the results of logistic regression. Studies showed that the Bayesian network in most cases demonstrated much better results that the classical classification method. The created decision support system showed high quality results, which indicates that the program is capable of detailed analysis of risks of losses.

The scientific novelty of obtained results is as follows:

— for the first time, selected types of generalized linear models for analysis of actuarial processes have been constructed and analyzed, which are distinguished by the possibility of their application for the formal description of linear and non-linear and non-stationary processes and provide the possibility of estimation of actuarial risk with high quality;

— for the first time, a new model based on LS-method in the form of a Bayesian network for the assessment of actuarial risks has been built, which is characterized by the correctness of the formal description of the selected process according to the Bayesian information criterion and provides the assessment of high-quality forecasts of actuarial processes;

— for the first time, a software-system methodology for building adaptive models of actuarial processes has been created, which differs in the proposed modeling in the presence of statistical uncertainties of the data, the method of estimating model parameters and provides an increase in the adequacy of the built models and quality of forecast estimates;

— the method for estimating the parameters of generalized linear models has been improved through the integrated application of the Monte-Carlo method for

Markov chains, the Adam method, and the iterative recursive weighted least squares method, which guarantees an increase in the quality of the parameters.

Keywords: nonlinear processes, Bayesian networks, LS-method, generalized linear models, logistic regression, iteratively-reweighted least squares method, Adam method, Monte-Carlo method for Markov chains, Kalman filter, simulation modeling, decision support system.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації:

1. Панібратов, Р. "Система підтримання прийняття рішень для оцінювання та прогнозування стану страхової компанії." *System research and information technologies,* no. 1, 2022, pp. 61-72. https://doi.org/10.20535/SRIT.2308-8893.2022.1.05 *(внесок: Панібратов Р.С. – розробка системи підтримки прийняття рішень, що дозволяє оцінювати та прогнозувати стан страхової компанії за її фінансово-економічними показниками за допомогою методів k-найближчих сусідів, методу опорних векторів, наївного байєсівського класифікатору, випадкового лісу, XGBoost та глибокої нейронної мережі з використанням даних страхових компаній України), фахове, Scopus.*

2. Panibratov, R., and P. Bidyuk. "Estimation of the Parameters of Generalized Linear Models in the Analysis of Actuarial Risks." *System research and information technologies*, no. 2, 2023, pp. 139-148. https://doi.org/10.20535/SRIT.2308-8893.2023.2.10 *(внесок: Панібратов Р.С. – реалізація запропонованих методів оцінювання параметрів узагальнених лінійних моделей; Бідюк П.І. – підготовка даних і уточнення узагальнених лінійних моделей), фахове, Scopus.*

3. Panibratov, R. S. "Analysis of data uncertainties in modeling and forecasting of actuarial processes." *Radio Electronics, Computer Science, Control*, no. 2, 2024, pp. 45-51. https://doi.org/10.15588/1607-3274-2024-2-5 *(внесок: Панібратов Р.С. – розробка і реалізація підходу для мінімізації статистичної невизначеності при аналізі та прогнозуванні актуарних процесів), фахове, Web of Science.*

4. Panibratov, R. S, and P. I. Bidyuk. “Analysis of actuarial risk with generalized linear models.” *System Research and Information Technologies*, no. 4, 2025, pp. 58-70. https://doi.org/10.20535/SRIT.2308-8893.2025.4.04 *(внесок:*

Панібратов Р.С. – проведення аналізу актуарних ризиків за допомогою узагальнених лінійних моделей з використанням методу Монте-Карло для марківських ланцюгів як для штучно згенерованих так і для фактичних даних; Бідюк П.І. – підготовка та аналіз даних для проведення досліджень), фахове, Scopus.

5. Panibratov, R., and P. Bidyuk. "Applying Bayesian networks in analysis of actuarial risks." *International Scientific Technical Journal" Problems of Control and Informatics"*, vol. 70, no. 3, 2025, pp. 33-44. https://doi.org/10.34229/1028-0979-2025-3-3 *(внесок: Панібратов Р.С. – побудова мережі Байєса, застосування методу auritzen-Spiegelharten для ймовірнісного висновку під час аналізу актуарних ризиків, порівняння результатів з розрахунками логістичної регресії; Бідюк П.І. – уточнення структури моделі), фахове, категорія Б.*

Наукові праці, які засвідчують апробацію матеріалів дисертації:

6. Panibratov, R. “Methods of estimating parameters of generalized linear models in the analysis of actuarial risks.” *Proceedings of V international scientific and methodical conference Computer technology and mechatronics*, Kharkiv, Ukraine, 24 May 2023, pp. 86-89. https://dl2022.khadi-kh.com/mod/resource/view.php?id=361231 *(внесок: Панібратов Р. С. – використання трьох нових методів оцінювання параметрів узагальнених лінійних моделей), вітчизняна конференція.*

7. Panibratov, R. "Analysis of data uncertainties in modeling and forecasting of actuarial processes." *Proceedings of 3rd International Scientific and Practical Internet Conference Scientific Research and Innovation*, Dnipro, Ukraine, 18-19 April 2024, edited by V. V. Marenichenko, WayScience, pp. 19-21. http://www.wayscience.com/wp-content/uploads/2024/04/Conference-Proceedings-April-18-19-2024-1.pdf *(внесок: Панібратов Р.С. – проведення аналізу невизначеностей в актуарних процесах та реалізація застосування методів фільтрації), вітчизняна конференція.*

8. Panibratov, R. “Analysis of actuarial risk with generalized linear models.” *Proceedings of 6th International Scientific and Practical Internet Conference Integration of Education, Science and Business in Modern Environment: Winter Debates*, Dnipro, Ukraine, 6-7 February 2025, edited by V. V. Marenichenko , WayScience, pp. 37-39. http://www.wayscience.com/wp-content/uploads/2025/02/Conference-Proceedings-February-6-7-2025-1.pdf *(внесок: Панібратов Р.С. – проведення аналізу актуарних ризиків за допомогою узагальнених лінійних моделей з використанням методу Монте-Карло для марківських ланцюгів як для штучно згенерованих так і для фактичних даних), вітчизняна конференція.*

9. Panibratov, R., V. Gavrylenko, and P. Bidyuk. “The system analysis based approach to market and credit risks analysis.” *Матеріали III Всеукраїнської наукової конференції здобувачів освіти і молодих учених Відбудова транспортної інфраструктури України*, Київ, Україна, 25 березня 2025 року, pp. 252-253. https://doi.org/10.33744/978-966-632-331-9-2025-1 *(внесок: Панібратов Р. С. – створення і представлення системи підтримки прийняття рішень для аналізу взаємодії ринкових та кредитних ризиків; Гавриленко В.В. – застосування моделі Васічека; Бідюк П.І. – застосування імітаційної моделі з використанням Монте-Карло), вітчизняна конференція.*

10. Panibratov, R. S. “Applying Bayesian networks in analysis of actuarial risks.” *Proceedings of 1st International Scientific and Practical Internet Conference “Crossroads of Ideas: Science, Technology and Society in a Global Context”*, Dnipro, Ukraine, 10-11 April 2025, edited by V. V. Marenichenko , WayScience, pp. 7-9. http://www.wayscience.com/wp-content/uploads/2025/04/Conference-Proceedings-April-10-11-2025.pdf *(внесок: Панібратов Р.С. – побудова мережі Байєса і застосування методу Lauritzen-Spiegelharten для ймовірнісного висновку під час аналізу актуарних ризиків), вітчизняна конференція.*

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

BIA — Basic Indicator Approach

LDA — Loss Distribution Approach

IMA — Internal Measurement Approach

Sb-AMA — Scenario-based Advanced Measurement Approach

ШІ — штучний інтелект

ERM — Enterprise Risk Management

LS — Lauritzen-Spigielharter

PAYD — Pay-As-You-Drive

УЛМ — узагальнені лінійні моделі

ANFIS — adaptive network-based fuzzy inference system

TSK — Takagi, Sugeno, Kang

RFE — recursive feature elimination

MCAR — missing completely at random

MAR — missing at random

MNAR — missing not at random

АРУГ — авторегресійна умовна гетероскедастичність

УАРУГ — узагальнена авторегресійна умовна гетероскедастичність

IRWLS — iteratively reweighted least squares

МНК — метод найменших квадратів

МКМЛ — метод Монте-Карло для марківських ланцюгів

RSS — residual sum of squares

AIC — Akaike information criterion

BIC — Bayesian information criterion

MSE — mean squared error

RMSE — root mean squared error

MAPE — mean absolute percentage error

СППР — система підтримки прийняття рішень

ВСТУП

Актуальність роботи. Актуарний ризик розглядається як ризик, коли реалізовані актуаріями припущення щодо моделювання цін страхових полісів можуть бути неточними або некоректними. Його рівень прямо пропорційний надійності припущень, реалізованих моделями ціноутворення, що використовуються страховими компаніями для встановлення премій.

Термін “страхування” є найбільш значущим чинником ризику та основою бізнесу як для страховиків так і перестраховиків. Але оскільки оформлення страхових полісів включає в себе широкий діапазон операційних, поведінкових та статистичних аспектів, різноманіття на набір ризиків є надзвичайно складними. Так як актуарії є доволі освіченими в питаннях розпізнавання ризиків, їх кількісного перетворення у змістовну інформацію та усвідомлення контексту та складності організації, вони є фахівцями в даній сфері. Розробка продукту, поведінка страхувальників, перестрахування, капітал, керування втратами та інші фактори мають бути взяті до уваги. Керівники страхових компаній покладаються на актуаріїв та інших професійних зацікавлених сторін для вказівок щодо визначення значущості ризиків. Досвід грає ключову роль у збереженні цілісності та довгострокової життєздатності компанії.

Оскільки ризики, що виникають, є новими та динамічними, вони створюють проблеми для актуарного моделювання. Щоб правильно виявляти та вимірювати ці ризики, актуарії мають модифікувати традиційні стратегії моделювання та створювати нові методи. Щоб актуарії могли пропонувати достовірні оцінки ризиків та сприяти прийняттю обґрунтованих рішень, мають бути вирішені проблеми моделювання. Оскільки ризики, що виникають, часто не містять вагомих історичних даних, то доволі складно використовувати звичайні актуарні підходи для прогнозування їх виникнення, складності та потенційного впливу. З метою доповнення обмежених історичних даних, актуаріям, можливо, потрібно покладатися на альтернативні джерела даних, як експертна думка, сценарний аналіз або зовнішні набори даних. Моделювання

сукупного впливу ризиків, що виникають, є по суті складною задачею, оскільки вони мають складні відношення та залежності з іншими ризиками. При моделюванні нових ризиків актуарії повинні враховувати можливість накопичення ризиків, ефектів поширення та нелінійних кореляцій. У зв'язку з браком даних та змінних чинників ризику, показники, що використовуються для моделювання ризиків, що виникають, як частота, серйозність та складність, часто є неточними. Для оцінки обґрунтованості припущень та вимірювання можливого діапазону наслідків, актуарії повинні включати аналіз чутливості та аналіз невизначеності у свої моделі. Різноманітні риси ризиків, що розвиваються, можуть бути недостатньо враховані традиційними актуарними моделями, такими як ті, що ґрунтуються на таблицях смертності чи досвіді попередніх страхових випадків. Для врахування динамічної характеристики, взаємозалежності, та невизначеності, що пов'язана з новими ризиками, актуарії повинні або змінити поточні моделі або створити нові структури моделювання.

Мета та задачі дослідження. Метою роботи дисертаційного дослідження є створення спеціалізованої системи підтримки прийняття рішень, що дає можливість детально аналізувати актуарні ризики, оцінювати можливі майбутні втрати та їх ймовірність.

Для досягнення поставленої мети необхідно розв'язати такі задачі:

— виконати аналіз існуючих методів оцінювання ризику, підходів та їх недоліки стосовно аналізу актуарних ризиків;

— побудувати моделі для оцінювання страхових ризиків у формі узагальнених лінійних моделей і Байєсівських мереж і логістичної регресії;

— створити систему підтримки прийняття рішень для оцінювання страхових ризиків;

— зібрати статистичні дані і виконати обчислювальні експерименти з метою оцінювання ризиків у страхуванні.

Об'єктом дослідження виступають фінансові процеси на страховому ринку, представлені статистичними даними.

Предметом дослідження є математичні моделі для опису і прогнозування процесів у страхуванні, методи оцінювання страхових ризиків.

Методи дослідження. При проведенні досліджень у межах дисертаційної роботи були використані такі методи:

— узагальнені лінійні моделі — використовувалися для кількісного оцінювання актуарного ризику;

— імітаційне моделювання — використовувалися для генерування страхових даних у зв’язку з малою доступністю інформації даного типу;

— метод оптимальної фільтрації даних Калмана — використовував при аналізі невизначеностей в актуарних процесах;

— Байєсівський аналіз даних — використовувався для ймовірнісного оцінювання актуарних ризиків;

— регресійне моделювання з використанням логістичної регресії — застосовувалося в якості класичних методів аналізу ризику.

Наукова новизна результатів, отриманих автором, полягає у наступному:

— вперше побудовано та досліджено вибрані типи узагальнених лінійних моделей для аналізу актуарних процесів, які відрізняються можливістю їх застосування для формального опису лінійних і нестаціонарних нелінійних процесів і забезпечують можливість оцінювання актуарного ризику з високою якістю.

— побудовано нову модель на основі методу Lauritzen-Spieigelharten у формі байєсівської мережі для оцінювання актуарних ризиків, яка характеризується коректністю формального опису вибраного процесу за байєсівським інформаційним критерієм і забезпечує оцінювання високоякісних прогнозів актуарних процесів.

— удосконалено метод оцінювання параметрів узагальнених лінійних моделей завдяки комплексному застосуванню методу Монте-Карло для марківських ланцюгів, методу адаптивного оцінювання моментів (Adam) та

рекурсивного ітеративно-зважувального методу найменших квадратів, що гарантує підвищення якості параметрів.

— створено і реалізовано програмно системну методологію побудови адаптивних моделей актуарних процесів, яка відрізняється запропонованим методом моделювання в умовах наявності статистичних невизначеностей даних, методом оцінювання параметрів моделей і забезпечує підвищення адекватності побудованих моделей та якості оцінок прогнозів.

Практичне значення отриманих результатів дисертаційного дослідження полягає у тому, що результати були використані у науковій та науково-дослідній роботі Інституту телекомунікацій і глобального інформаційного простору національної академії наук України для розв’язання задач дослідження ризиків, розроблення методик їх оцінювання та визначення потенційних наслідків.

Особистий внесок здобувача. Всі основні результати дисертаційної роботи отримані автором особисто.

Всі 5 фахових публікацій написано у співавторстві, а здобувачеві належать такі результати. В роботі [1] здобувачем розроблено систему підтримки прийняття рішень, що дає можливість оцінювати та прогнозувати стан страхової компанії за її фінансово-економічними показниками за допомогою методів *k*-найближчих сусідів, методу опорних векторів, наївного байєсівського класифікатору, випадкового лісу, XGBoost та глибокої нейронної мережі; надано результати прогнозування майбутнього стану страхових компаній України. В роботі [2] здобувачем було запропоновано застосувати три методи оцінювання параметрів узагальнених лінійних моделей, що являло собою рекурентний ітераційно-зважувальний метод найменших квадратів, метод Adam і метод Монте-Карло для марківських ланцюгів; виконано порівняльний аналіз методів на основі метрик якості прогнозу. В роботі [3] здобувачем було проведено аналіз невизначеностей в актуарних процесах та запропоновано застосування методів фільтрації; проведено обчислювальний експеримент для трьох зазначених вище алгоритмів оцінювання параметрів із

та без застосування оптимального фільтру Калмана і зроблено висновки щодо запропонованого підходу. В роботі [4] здобувачем було проведено аналіз актуарних ризиків за допомогою узагальнених лінійних моделей з використанням методу Монте-Карло для марківських ланцюгів як для штучно згенерованих так і для фактичних даних; виконано порівняльний аналіз запропонованих моделей на основі критеріїв адекватності моделей. В роботі [5] здобувачем побудована модель у формі мережі Байєса і застосовано метод Lauritzen-Spiegelharten для ймовірнісного висновку під час аналізу актуарних ризиків; результати були порівняні з розрахунками логістичної регресії, як класичного методу класифікації.

Апробація результатів дисертації. Результати та основні положення роботи доповідалися та обговорювалися на наступних конференціях: The Fifth International Scientific and methodical Conference “Computer Technology And Mechatronics” (Kharkiv, Ukraine 24.05.2023) [6]; The 3rd International Scientific and Practical Internet Conference “Scientific Research and Innovation” (Dnipro, Ukraine, 18-19.04.2024) [7]; The 6th International Scientific and Practical Internet Conference “Integration of Education, Science and Business in Modern Environment: Winter Debates” (Dnipro, Ukraine, 6-7.02.2025) [8]; III Всеукраїнська наукова конференція здобувачів освіти і молодих вчених “Відбудова транспортної інфраструктури України” (Київ, Україна, 25.03.2025) [9]; First International Scientific and Practical Internet Conference “Crossroads of Ideas: Science, Technology and Society in a Global Context” (Dnipro, Ukraine, 10-11.04.2025) [10].

Публікації. За матеріалами дисертації опубліковано 10 робіт, з яких п’ять — це статті у журналах і збірниках наукових праць, що входять до переліку фахових видань затверджених МОН України за спеціальністю дисертації (1 включена до міжнародної наукометричної бази Web of Science, 3 включені до міжнародної наукометричної бази Scopus), та 5 публікацій у матеріалах конференцій.

Структура та обсяг дисертації. Дисертація складається із анотації, вступу, чотирьох розділів, висновків, списку використаних джерел, додатків.

Робота містить 198 сторінок, у тому числі: 172 сторінок основного тексту, 47 рисунків, 18 таблиць, список використаних джерел із 73 найменувань на 9 сторінках.

РОЗДІЛ 1 РИЗИКИ У СТРАХУВАННІ ТА ЇХ МЕТОДИ ОЦІНЮВАННЯ

Фінансові ризики мають зростаючий та значущий вплив на ефективність та рівень фінансової стабільності різних видів бізнесу, які залучені в різноманітні види діяльності. У зв’язку з економічною нестабільністю країни, розробкою новітніх та креативних фінансових інструментів, збільшення масштабів фінансових відносин, динамічністю ситуації на фінансовому ринку та інших чинників, фінансові ризики для бізнес-осіб мають серйозний вплив для наслідків економічних дій.

Надійність страховиків, які є серйозними гравцями на ринку, являє собою першочергову вимогу для належного функціонування страхових ринків. Підтримка вміння кожного страховика відповідати своїм зобов’язанням в повному обсязі та вчасно, тобто говорячи про фінансову стабільність, є відправною точкою для реалізації функцій страхування. На сьогодні фінансова ситуація страхових компаній потребує пошуку нових засобів з метою покращення конкурентоспроможності та фінансової надійності. Звідси виникає необхідність у розробці систем для більш ефективного оцінювання фінансової ситуації та стійкості страхових компаній.

Детальне дослідження ризиків з наданням економічних та математичних аргументів щодо фінансової діяльності страхових компаній є основною вимогою у зв’язку з діапазоном засобів проявів ризику разом з частотою та ступеню наслідків застосування. За рахунок здатності надання дійсних та надійних результатів з метою аналізу та керування фінансовими ризиками використовуються математичні та економічні підходи. Ймовірність настання банкрутства, власний капітал, маржа платоспроможності є головними індикаторами для свідчення щодо фінансової стійкості страхових компаній.

Тому до найбільш важливих зобов’язань в контексті постійних операцій різних типів бізнесів, включаючи страхові компанії, відносять ідентифікацію, оцінювання, та управління ступенем фінансового ризику.

1.1 Стан страхового ринку України та світу

Одним із значущих складових фінансової системи будь-якої країни є страховий сектор. Він захищає фізичних та юридичних осіб від грошових втрат, що спричинені внаслідок різноманітних причин. За рахунок зниження ризиків та заохочування інвестицій та економічного зросту, зріст страхової сфери допомагає вдосконалювати економічну безпеку держави. Можливості та задачі ринку серйозно збільшилися разом з пропозицією стійкого захисту проти неочікуваних фінансових ризиків та допомогою розвитку надійних економічних обставин. Страховий ринок — галузь бізнес-угод, де об’єктом купівлі та продажу є страхове покриття. Послуга, що компенсує страховику втрати, що виникли для застрахованої особи у зв’язку з виникнення страхової події носить назву страховий захист.

Структура страхового ринку зображена на рисунку 1.1 [11].

Страховий ринок функціонує наступним чином: страхувальник укладає страховий договір зі страховою компанією; клієнт виплачує компанії страхову премію, і в разі настання страхової події, страховик отримує грошову компенсацію. З метою гарантії фінансової стабільності страхові компанії відіграють значущу роль. Вони накопичують гроші шляхом збору премій, які згодом використовують для покриття страхових відшкодувань у випадку подій, згідно укладеному договору. Завдяки цьому засобу, страхувальники мають можливість знизити власні фінансові ризики, тим самим покращуючи надійність економіки.

Страхові рішення дають можливість захистити підприємства, типи власності і фізичних осіб від шкоди, завданої стихійними лихами та різними правопорушеннями. Завдяки захищеності від можливих фінансових збитків, люди та компанії мають здатність працювати та інвестувати з повною впевненістю.

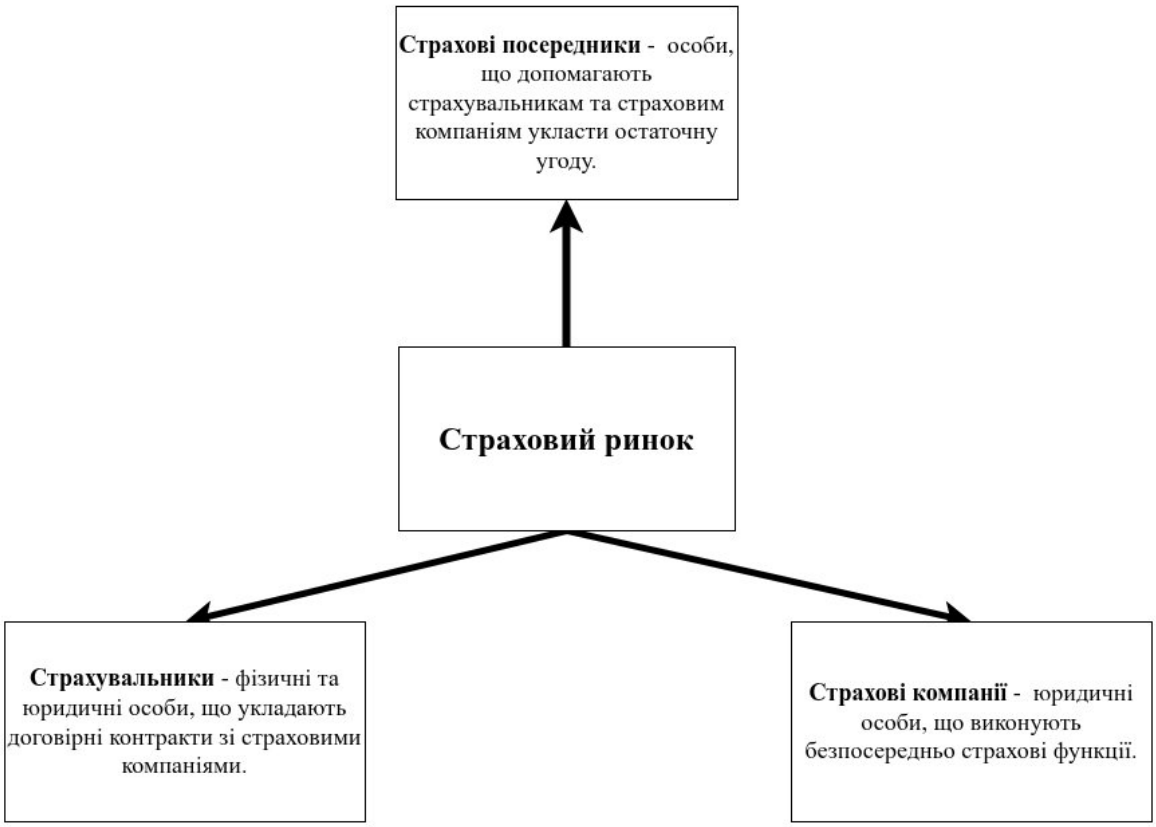

Рисунок 1.1 — Структура страхового ринку

Станом економічної безпеки називають умову, за якої економіка, суспільство та його громадяни захищені від зовнішніх та внутрішніх факторів небезпеки, що можуть ставити під загрозу стабільність, зростання та економічну активність.

Здатність запобігати, захищатись та підлаштовуватись до різних видів ризиків у вигляді інфляцій, природних та техногенних катастроф є важливою складовою економічної безпеки. Це включає складові соціальної, економічної та фінансової стійкості і загальну здатність суспільства витримувати діапазон внутрішніх та зовнішніх стресів. Економічна безпека впливає на соціальне та економічне зростання і є важливою частиною безпеки загалом. Реалізація політики та планів вдосконалення економічної безпеки допомагає стабілізувати економіку, покращити довіру суспільства до фінансових установ і якість життя

громадян. Як наслідок, зв’язок між безпекою та страхуванням допомагає покращити економічну безпеку держави.

Зв’язок страхового ринку з економічною безпекою виражається такими напрямами, представленими нижче.

1. **Захист майна і компенсація збитків.** Компанії та власники нерухомості можуть захистити власні активи від різноманітних небезпек завдяки послугам страхових компаній. Це мінімізує фінансові збитки та позитивно впливає на довгострокову економічну тривалість.

2. **Гарантія фінансової надійності.** Страхові організації грають роль у наданні послуг для фізичних так і юридичних осіб у керуванні ризиками та неочікуваними обставинами. Це робить можливим гарантування фінансової стійкості звичайним людям та підприємствам.

3. **Страхування життя.** У випадку смерті годувальника родини або його тяжкої хвороби, страхування життя або медичне страхування забезпечує підтримку страхувальника та його сім’ї. Це вдосконалює економічну активність та допомагає в управлінні соціальної стабільності.

4. **Захист від відповідальності.** Страхування професійної та цивільної відповідальності дає захист від грошової шкоди внаслідок судових справ та вироків.

5. **Профілактика безпеки та їх заохочення.** Різні страхові установи надають власним клієнтам інструкції та рекомендації щодо запобігання ризику. Це підвищує рівень інформованості щодо проблем безпеки та допомагає уникнути потенційних небезпек.

6. **Правила безпеки та інструкції.** Страхувальники мають уповноваження щодо встановлення інструкцій безпеки та їх стандарти для власних клієнтів.

7. **Фінансування діяльності та проектів, що пов’язані з безпекою.** Страхові компанії можуть забезпечувати фінансову підтримку програм та проєктів, пов’язаних з безпекою, в різних сферах, що включає будівництво техніку безпеки з транспортування.

Уміння сучасної страхової сфери адаптуватися до соціальних змін та проблем сьогодення є одним із основних функцій. Це охоплює вирішування категорій ризику, що з'являються, як кібербезпека та запуск страхових планів, що позитивно вливають на екологічну відповідальність та сталий розвиток. Велика кількість проблем та подій в суспільстві мають вплив на сферу страхування. Це відомо як цифрове перетворення, і це включає зростаюче застосування методів, як аналіз даних та мобільні додатки, що впливає на процес продажу та обслуговування страхових продуктів. Крім того фіксуються зміни клімату, що збільшує ризики для страхових організацій, особливо в галузі нерухомості та стихійних лих. Демографічні зміни та медичні технології також мають вагомий вплив на сферу страхування. Довгострокові страхові поліси та поліси страхування здоров'я можуть бути під впливом змін чисельності населення та введення нових технологій медицини [12].

Огляд страхування України показав, що, незважаючи на кризу, ця галузь продовжує працювати та поступово покращується. Зниження обороту клієнтів, зменшення у спрощеннях регуляторних вимог, зменшення об'ємів роздрібних послуг страхування, поширеність трендів відкладених платежів серед клієнтів страховиків, недостатня довіра з боку громадськості, відсутність гарантій у разі розорення страхової організації, нестача прав для споживачів послуг страхування, брак привабливих можливостей в інвестуванні для страхових компаній та нестача керування страховими організаціями є основними проблемами, що визначають страховий ринок станом на сьогодні [13].

Динаміка страхового ринку, збільшення обсягів страхових операцій, забезпечення страхових послуг, гнучкість та здатність пристосовуватися до вимог сучасності і інтерес іноземних інвесторів вказує на те, що, не дивлячись на численні проблеми та завад, з якими зіштовхується українська страхова сфера, є можливості для постійного зросту та інтеграцію у глобальну арену. З метою гарантії відновлення та зростання ринку страхування, необхідно [14]:

1) перевірити та оновити страхові закони згідно дійсних стандартів;
2) дати опис стратегії страхової сфери та вектору розвитку;

3) встановити вимогу для платоспроможності страховиків;

4) провести огляд умов видачі ліцензій та причин призупинення та анулювання;

5) активізувати зусилля маркетингу, провести оновлення робочої сили та розробити абсолютно нові підходи до продажу страхових продуктів;

6) почати навчання суспільства стосовно культури страхування та надати найбільш можливу підтримку на рівні держави.

Європейські стандарти і технології мають потенціал модернізувати український страховий бізнес, покращують конкурентоспроможність та пропонують великий набір можливостей. Кількість страхових компаній, відсоток премій страхування життя та об'єм валових страхових виплат можуть значно зрости в Україні. У порівнянні з ЄС, Україна має низький рівень проникнення страхування, що дає простір для росту у сфері страхування. Але щільність страхування в Україні значно нижча, ніж в більшості країн ЄС, що підкреслює потребу в розвитку та вдосконалення страхових послуг [15].

Вітчизняні страховики станом на сьогодні мають не тільки збільшувати кількість страхових послуг, що вони пропонують у більш складні економічні періоди, але крім того і диверсифікувати власні пропозиції щодо продукту, щоб запропонувати власним клієнтам кращі страхові продукти, що можуть їх захистити одночасно від нових і традиційних ризиків, принесені війною. Оскільки їх впровадження дає суспільству варіанти накопичення капіталу та довгостроковий страховий захист разом з додатковими інвестиційними ресурсами для економіки держави, задача збільшення доступності послуг страхування життя в Україні лишається актуальною [16].

Епідемія COVID-19 стала причиною суттєвих змін у сфері страхування як України та і по всьому світі. Страхування стрімко розвивалося і були зроблені прогнози оптимістичного характеру щодо 2020 року. Але COVID-19 спричинив значний спад в глобальній економіці, що однозначно вплинуло на сферу страхування. Були як переваги, так і недоліки для нових обставин роботи. Оскільки страхування власності не було пріоритетним через карантин, пандемія

мала серйозний негативний ефект на нього. Автострахування також було під впливом втрат, особливо внаслідок менших обсягів нових страхових контрактів. З іншого боку, аналіз динаміки зросту ринку страхування України продемонстрував, що пандемія мала протилежний ефект в сенсі заохочування розвитку страхової культури в країні та переконання в плані цінності страхової послуги, особливо, коли йде мова про страхування життя та здоров'я. Зростання цифровізації та інновацій у сфері страхування стало ще одним позитивним наслідком [17].

Застосування певного набору технологій та системи телекомунікації з метою заохочування вдосконалення якості обслуговування потенційних так й існуючих клієнтів називається цифровізацією. Воно розглядається як новий засіб створення та розвитку страхового ринку. Акцент на перетворенні суспільства у нове, де операції ґрунтуватимуться на використанні різноманітних цифрових технологій є одним з чинників, що позитивно впливає на подальше вдосконалення цифровізації страхового ринку. До основних тенденцій цифровізації, властивих страховому ринку України можна віднести наступне [18]:

— масштабування цифрової страхової системи, що представляє собою комплекс різноманітних технологій, що пропонується страховиками;

— формування мобільного страхового механізму за рахунок створення відповідних спеціалізованих мобільних додатків;

— заохочування використання хмарних технологій, що гарантує збереження даних страхової компанії;

— використання страховиками автоматизованих системи верифікації контрактів протягом операції, що водночас дозволяє уникнути негативного впливу шахрайських дій та вдосконалює поточну систему обслуговування клієнтів страхової організації;

— пропозиція застосування телематики у сфері страхування: завдяки цьому підходу страхові компанії здатні встановити найбільш ефективну та ідеальну страхову премію.

Страхова сфера України зіткнулась з серйозними перешкодами під час воєнного стану, але водночас змогла вистояти. Передбачається, що більш жорсткий контроль та регуляція змогли б підтримати відновлення страхового ринку та майбутньому зростанню. Покриття збитків внаслідок військової діяльності є першочерговою задачею, з яким зіштовхується страховий сектор під час воєнного стану. Перегляд ставлення страхових компаній до можливостей покриття ризику є важливим разом з адаптацією умов страхового продукту до умов воєнного стану, як врахування ризиків, що пов’язані з безпосереднім впливом військових операцій та їх наслідками [19].

Детальний огляд динаміки, переліку страхових компаній, премій та виплат дав можливість отримати інформацію щодо впливу військового конфлікту на ринок страхування України. Після першочергового шоку, основним напрямом була адаптація та стабілізація функцій страхових компаній у нинішній час. У період ринкової нестійкості, співвідношення страхових премій до виплат демонструє ступінь стабільності ринку як страхові компанії намагаються відповідати фінансовим зобов’язанням. Зміни в попиті та поширеності різних типів страхування серед фізичних та юридичних осіб були продемонстровані завдяки аналізу структури страхових премій за видами страхування. Це важливо для подальшого зростання ринку. Аналіз страхових активів протягом кризового періоду, що досліджувався, показав фінансово-життєвий стан організацій та дав інформацію про їх можливість долати більше, ніж економічні випробування. Згадані проблеми у розвитку страхової сфери України дає роздуми щодо можливих підходів з метою реорганізації та вдосконалення фінансового середовища за рахунок введення нових планів та методів [20].

Активи страховиків ризику зросли на 3% та 11% відповідно, у той час як активи страховиків життя зросли на 5% у четвертому кварталі та на 14; у 2024 році. 10 найбільших ризиків станом на зараз нараховують 71% об’єму премій у порівняння з минулим 65%. Майже 50% премій страхування життя оплачується найбільшим страховиком. Хоча страхові премії страховиків ризику сезонно

зменшилися до всього 1% у четвертому кварталі, щорічний приріст був усього 15%. Кількість сплачених страхових премій була вищою, ніж у 2021 році. Ріст премій для страхування життя прискорився до 17% у четвертому кварталі та 5% за весь рік. У 2024 році страховики ризиків отримали чистий прибуток обсягом 2,5 млрд гривень, що є на 31% більше у порівняння з попереднім роком. Рентабельність капіталу зросла до 14%, що на 4 відсотки вище. У зв'язку з фінансовими результатами одного великого страховика, прибуток страховиків життя за рік зріс більше ніж у 4 рази до 1,4 млрд гривень. Водночас рентабельність капіталу лишилася майже незмінною. Усі страховики виконали критерії мінімального капіталу (MCR) та капіталу платоспроможності (SCR) до кінця 2024 року. Показник SCR для трьох бізнесів знаходиться в діапазоні між 100 та 120 відсотків [21].

Внаслідок змін у сфері страхування життя з'явилося різноманіття можливих проблем, пов'язаних з фінансовою стабільністю, що вказує на зміну профілю ризику галузі та зростання взаємозалежності між ширшою фінансовою системою. Зростання вразливості до ризику, менш ліквідних активів, складність, викликана угодами перестрахування з інтенсивним використанням активів, та концентровані ризики в певних країнах та компаніях є наслідками змін, що сприяли зростанню зменшенню потреб у капіталі. З метою обмеження можливих ризиків та обслуговування важливу позицію сектору в фінансовій системі, дані події потребують збалансованої та агресивної відповіді урядів [22].

Зміна структури сектору значно ускладнила для наглядових та регуляторних органів належний моніторинг та керування ризиками. Впровадження власних алгоритмів оцінювання для неліквідних активів та зростаюча залежність від приватних ринків, де вартість зазвичай є непрозорою, маскує профілі ризиків за стає на заваді до прозорості. Оскільки офшорні угоди про перестрахування приховують повну картину ризиків та веде до юрисдикційних невідповідностей, вони роблять задачу управління ризику та оцінку платоспроможності більш складною. Ці вразливості, тобто великі

інвестиції в непрозорі активи, збої офшорних перестрахувальників та подальше повернення зобов’язань до цедентів, що може призвести до проблем ліквідності та примусового продажу активів, підкреслюються банкрутством 777 Re у 2024 році, бермудського перестраховика, пов’язаного з приватним акціонерним капіталом [22]. Також керівники та зацікавлені сторони не здатні розпізнавати накопичення небезпек у зв’язку з недостатньою доступністю даних.

Небезпека синтетичного левериджу через деривативи може також зростати завдяки структурним змінам зі сторони активів. Наприклад, зростання синтетичного левериджу через ф’ючерси обмінного курсу відображає зростаючий вплив страховиків, пов’язаних з приватним сектором, на активи, деноміновані в валютах, відмінних від долара США. Для збереження грошей га вимогах до застави, страховик явно відходять від центральних контрагентів та переходять до двостороннього клірингу свої деривативних контрактів. Внаслідок цього, страховики особливо занепокоєні по відношенню до ризиків, пов’язаними з двостороннім клірингом позабіржових деривативів [22].

1.2 Поняття про ризик у галузі страхування

Говорячи про ризик, йдеться мова про можливу небезпеку втрат внаслідок певних стихійних лих, випадкових ситуацій техногенного характеру, злочинних дій та різноманітних економічних явищ. В останньому випадку види цього типу стрімко зростають внаслідок вдосконалення кредитної системи та серйозної зміни поділу праці. Страховий ризик та страхові дії нерозривно пов’язані між собою. Водночас розподіл ризику є базовим елементом сфери страхування. Першочерговою умовою для створення страхових відносин є ризик. За відсутності останнього зникає необхідність в укладанні договору. Водночас потрібно звернути увагу, що далеко не всі типи ризику підпадають під категорію для створення відносин у страхування. Тільки ті види, які придатні до оцінювання ймовірності настання страхової події та обчислення можливих втрат і рівномірних страхових премій відносять до множини ризиків,

здатних до покриття. Під терміном страховий ризик як правило досліджують очікувані втрати об’єкту страхування у випадку виникнення страхової події [23].

Оскільки керування ризиком вимагає передусім здатності аналізувати та оцінювати його, то оцінювання ризиків у фінансах є одним з першочергових етапів фінансового аналізу. Кількісне оцінювання факторів та видів ризику відоме як оцінювання ризику. Інакше кажучи, метод аналізу ризику включає додаткові кількісні та якісні методи.

Майже у всіх теоретичних дослідженнях щодо проблем невизначеності робиться акцент на класифікації, причини появи та кількісні методи оцінювання ризиків. З іншого боку, підходам якісного характеру, що включають в себе вплив чинники не-економічного походження на діяльність економічного та фінансового характеру приділяють значно менше уваги. Ці методи були б набагато кращими для дійсних можливостей вітчизняних підприємців в реальності, а також вони б слугували підґрунтям для майбутніх досліджень.

Звідси зростає необхідність у застосуванні економічних та математичних методів моделювання та аналізу фінансових ризиків та у створенні актуарних моделей з метою пошуку оптимального значення страхових премій за обчисленою ймовірністю банкрутства. Це пов’язано з тим, що знаходження ймовірності розорення страхової компанії є однією з найважливіших задач страхової математики, де головні актуарні ідеї фінансової оцінки стійкості створюються, беручи до уваги не лише відсутність банкрутства, але також і його запобігання.

Здатність висувати припущення по відношенню до розподілу страхових виплат та часових інтервалів між виплатами, що виконує базову роль для побудови моделі аргументує використання актуарних моделей. Так, процес Пуассона може застосовуватися для опису послідовностей виплат. Виплати можуть мати один і той же розподіл з апріорі відомою чи довільною функцією розподілу. Крім того, інтервали між виплатами можуть містити нерівномірні

параметри розподілу і можливі інші ситуації під час виплати дивідендів учасникам.

1.3 Методи оцінювання ризиків

Оцінка ймовірності виникнення події втрат та можливий обсяг втрат в разі настання є двома чинниками, що встановлюють ступінь ризику. Важливо акцентувати на тому, що можливі втрати як правило мають непередбачуваний характер. Для оцінки ймовірності настання даної події використовують як об'єктивний, так і суб'єктивний підхід. Обчислення частоти події зазвичай застосовують в якості об'єктивних методів. Суб'єктивний критерій, що будується на множині припущень формують базис суб'єктивного методу. Звідси встановлення кореляції між спрогнозованими розмірами втрат та ймовірністю їх настання являє собою суттю оцінювання фінансового ризику. Даний зв'язок виникає у формі ймовірнісної кривої наперед заданого ступеня ризику або настання втрат.

Ефективні математичні моделі наразі є в доступними для оцінювання фінансових ризиків, що мають відношення до страхування. Так, створення моделей керування та оцінювання ризиків у сфері фінансів підтримується найбільш популярними стандартами Basel II та Solvency.

В страховій галузі, ризик оцінюється шляхом застосування безліч унікальних підходів, що постійно еволюціонують та покращуються. Звідси розрізняють два алгоритми для здійснення аналізу ризику: дослідження факторів ризику та дослідження наслідків виникнення ризику. Даний поділ групує математичні моделі для оцінювання ризику у два головних класи: високорівневі моделі, що ґрунтуються на аналізі результатів, та низькорівневі моделі, що ґрунтуються на дослідженні чинників ризику. В якості критерію класифікації виступає набір задач, що дана модель здатна вирішити. Такі величини як середні втрати страхової фірми та максимальні втрати з апріорі визначеним рівнем значущості оцінюються шляхом використання моделей, що

підтримують явну функцію розподілу збитків.

Інструменти математичної статистики виступають в ролі бази для моделей, що аналізують кінцевий результат. Основна частина підходів включає збір та аналіз даних щодо фінансових втрат, що мають відношення до ризику протягом минулих періодів з подальшою екстраполяцією на наступні. До них відносять такі методи [24-27].

BIA: Стандарти, що гарантують достатній об’єм капіталу для вирішення проблем ризиків окреслюються в моделі BIA (Basic Indicator Approach) або в методі базових показників. За рахунок даного підходу вирішується задача обчислення максимального об’єму втрат за умови певного рівня значущості. Передбачається гіпотеза, що операційні втрати є побічним результатом усіх збитків, що пов’язані з ризиком.

Наступна формула визначає квантиль розподілу випадкової змінної x, що описує обсяг збитків:

$$\tilde{Q}_{99\%} = \hat{G}_I \alpha, \tag{1.1}$$

де $\tilde{Q}_{99\%}$ виступає в якості оцінки 99%-го квантиля втрат в якості випадкової величини;

$\hat{G}_I$ — середнє значення валового доходу за минулий трьохрічний період, що характеризував позитивну динаміку;

α = 15% — коефіцієнт, встановлений Базельським Комітетом згідно аналізу практики банків та фінансових інституцій.

Результати оцінювання збитків, що пов’язані з фінансовим ризиком страхових компаній України зазвичай застосовуються з метою коригування значення даного коефіцієнта. Мінус моделі BIA полягає в тому, що оцінка обсягів можливих фінансових втрат ґрунтується лише на обсягах бізнесу і не має зв’язок з властивостями страхового портфелю.

LDA. Передумовою методу LDA (Loss Distribution Approach) або методу розподілу втрат є те, що випадкова змінна x, що описує обсяги збитків виникли протягом періоду k можна описати наступним чином:

$$x = \sum_{j=1}^{n(k)} L_j, \tag{1.2}$$

де $n(k)$ — випадкова величина, що позначає кількість випадків збитків певного типу за визначений часовий інтервал k;

L_j — випадкові величини, що характеризують втрати за певний період.

Для певного типу втрат, висувається гіпотеза щодо незалежності то одного й того ж розподілу значень множини L_j Для кожної пари формату "галузь бізнесу-тип втрат", виконується аналіз збитків за певний часовий проміжок з метою створення даних моделей. Значення вибіркового середнього частоти виникнення ризикових випадків $E(n(k))$ та значення вибіркового середнього від обсягу збитків у разі настання ризикової події $E(L)$ обчислюються для кожної пари згідно даних щодо частот втрат та обсягу, що спостерігалися за попередній період. Функція розподілу випадкової змінної x генерується методом Монте-Карло після визначення функцій розподілу випадкових змін $n(k)$ та L. Квантиль заздалегідь встановленого рівня або значення операційного значення ризику або OpVaR можна розрахувати, точкову оцінку математичного сподівання втрат може бути обчислено за допомогою застосування функції розподілу випадкової величини x, що описує повний об'єм збитків для певного ризику. Варто акцентувати на тому, що окремі реалізації методів класифікують за двома групами. Не враховуючи фактори ризику та причинно-наслідкових взаємодій в межах контексту більш всеохоплювальної моделі, перша категорія моделі ґрунтується на прямих перевірках виникнення ризику та пов'язаних з ним збитків. Дані алгоритми охоплюють підходи, що базуються на теорії екстремальних значень та пряме

оцінювання ознак частотних розподілів та дій виникнення ризику. Більш широкий діапазон змінних, що включає чинники ризику, утворює базис другого набору методів [28].

IMA: Без створення розподілу випадкової змінної x, що описує обсяг втрат, метод IMA (Internal Measurement Approach) або метод внутрішніх показників дає можливість оцінити максимально ймовірні збитки для певного типу ризику. Існує функціональна залежність між значення очікуваних та неочікуваних збитків, якщо розділити усі витрати на дві категорії: непередбачувані втрати, тобто ті, що перевищують середнє значення та потрапляють у хвостову область статистичного розподілу, та очікувані втрати, що є сумою втрат, які близькі до математичного сподівання за певний період.

Лінійна залежність є прикладом базової ситуації:

$$\tilde{Q}_{99\%} = E_I P_E L_{GE} = \gamma E_L \alpha, \tag{1.3}$$

де $\tilde{Q}_{99\%}$ — оцінка 99% квантиля розподілу можливих збитків за певним типом події, тобто обсяг капіталу для покриття ризику;

P_E — ймовірність виникнення певної події за період часу;

L_{GE} — середній обсяг втрат у разі настання негативної події;

E_I — коефіцієнт масштабування;

γ — коефіцієнт оцінювання обов'язків щодо капіталу через оцінювані збитки E_L.

Сума всіх оцінок, розрахована для всіх видів ризикових подій та бізнес-галузей є остаточною оцінкою найбільш можливих втрат.

Статистичний метод: Даний підхід передбачає аналіз статистик втрат, які являють собою несприятливі наслідки здійснених операцій, що відбулися в інших порівняних економічних видах діяльності. Дисперсійний аналіз є одним із алгоритмів оцінювання, що може бути використаний в даному випадку. Частота збитків, яка має зв'язок з певним видом діяльності є першочерговим

критерієм, що визначається шляхом застосування наступного статистичного методу:

$$f = \frac{\widehat{m}}{m_{total}}, \tag{1.4}$$

де $\widehat{m}$ — число подій, в яких виникли втрати внаслідок несприятливих подій, у вибірці;

m_{total} — загальне число випадків з, що аналізувалося.

Показник частості збитків застосовується з метою прогнозування даних і береться до уваги в якості ймовірності певного ступеня збитків під час прийняття рішення при використанні даного підходу.

Статистичний підхід до моделювання, який має назву метод Монте-Карло, являє собою широко використовуваний варіант статистичного алгоритму станом на сьогодні. Перевагою підходу є те, що він дозволяє аналіз та оцінювання багатьох сценаріїв розробки проекту з урахуванням різноманіття елементів впливу в межах однієї структури. Значна кількість ймовірнісних функцій, що застосовується у даній стратегії є недоліком, який відштовхує менеджерів проєктів.

Крім того є можливість кількісно оцінити вплив якісно однорідних явищ по відношенню середнього рівня розвитку та представити типові розміри різних аспектів під час застосування середніх значень. Одним із статистичних засобів дослідження є підхід на основі оцінювання середнього значення. З метою надання аналітичної бази для оцінювання ризиків за певними характеристиками ризику, припускається, що індивідуальні групи ризику діляться на менші підгрупи під час процесу оцінювання. Конкретний набір позитивних та негативних відхилень від середнього типу ризику даної аналітичної основи представляється в рамках оцінювання ризику за методом відсотків, що являє собою метод статистичного аналізу. Ці відхилення від типового виду ризику вимріюються у відсотках.

Також застосовуються статистичні та економетричні алгоритми оцінювання та аналізу ризику. Слід звернути увагу на те, що вітчизняні та міжнародні дослідники мають надійний інструментарій для оцінки та моніторингу ризику у сучасному світі. Дані інструменти включають актуарні розрахунки, моніторинг, моделювання внутрішніх процесів у страхуванні, емпіричні дослідження, засоби експертного оцінювання, асоціації та аналогії, тощо.

Моделі на основі аналізу факторів ризику дають можливість застосовувати дані щодо внутрішніх причинно-наслідкових зв'язків та пропонуються детальний огляд процесів організацій. До цієї групи відносять наступні методи.

Sb-AMA: Метод Sb-AMA (Scenario-based Advanced Measurement Approach) або сценарний аналіз ґрунтується на ідентифікації величин ризику або можливих джерел ризику, з яких будуються сценарії формату "що буде, якщо станеться наступне" [29]. Звідси, дана модель створюється на основі втрат, що можуть виникнути у майбутньому у разі настання події, у противагу до зазначених методів вище, що включають в себе дослідження втрат, що вже відбулися. Експертні оцінки або історичні дані застосовуються з метою розрахунку частоти та обсягу збитків для кожного сценарію. Сценарії поділяються за змінними ризику, коли дані пройшли перевірку і неточні оцінки були скориговані. Параметри статистичних розподілів частоти та об'єму втрат оцінюються для кожної групи сценаріїв.

Оцінка повного розподілу втрат є здійсненою за допомогою методу Монте-Карло для імітаційного моделювання, що представляє собою останній крок. Оцінка виконується за умови припущення, що розподіл випадкових величин розміру втрат має форму, як розподіл Пуассона, та обсяг втрат, що відповідає заданому сімейству. В останньому випадку зазвичай використовуються закони розподілу з важкими хвостами. Звідси, постійне спостереження фінансового стану знижує ризик розорення підприємства за рахунок раннього виявлення несприятливих тенденцій.

Метод функціональних компонент. Основою даного підходу є розробка організаційної структурної моделі [30,31]. У цьому випадку, зсув до математичної моделі виконується за рахунок побудови орієнтованого графу з інформаційними потоками в якості його дуг та працівників і відділів в якості вершин. Випадковий процес присвоюється кожній функціональній вершині, що відповідає їй. Даний підхід є унікальним у тому, що він встановлює стохастичні зв'язки між функціональностями пов'язаних вершин. У даному випадку аналіз ризику виконується по відношенню до співпрацівників, систем та бізнес процесів, яким призначені вершини графу, аніж з точки зору наслідків конкретних чинників ризику.

Подібно до методу LDA, для побудови розподілу значень втрат для вершини i застосовується випадкова величина L_i, що має зв'язок з конкретним випадком реалізації ризику. Випадковий процес $n_i(t)$, що має два стани $n_i(t) = 0$, який позначає нормальне функціонування та $n_i(t) = 1$, який позначає відмову, описує випадки реалізації ризику для вершини i. Несправності для вершини i у графі призводять до остаточних втрат $X_i(t + \Delta t)$ у момент часу $(t + \Delta t)$, що можна записати наступним чином:

$$X_i(t + \Delta t) = X_i(t) + n_i(t + \Delta t)L_{i,t+\Delta t}, \tag{1.5}$$

де $L_{i,t+\Delta t}$ — реалізація випадкової величини L_i.

Концепція "підтримки", яку кожен функціональний елемент отримує від елементів, що пов'язані з ним, надається з метою визначення залежності між різними функціональними елементами. Зв'язаний функціональний елемент переходить у стан відмови, тобто $n_i(t) = 0$, якщо рівень підтримки знаходиться нижче заданого порогового значення.

Випадкові процеси $h_i(t)$ вводяться з метою представлення структури:

$$h_i(t) = \xi_i - \sum_j w_{ij} n_j(t) + \varepsilon_i(t), \quad (1.6)$$

де ξ_i — середній об'єм "підтримки", яку процес i отримує за умови повного функціонування системи, коли усі $n_i(t) = 1$;

w_{ij} — об'єм підтримки переданий від процесу j до процесу i.

$\varepsilon_i(t)$ — Гаусів білий шум.

Даний підхід робить акцент на аналізі функціональності процесів та виявлення недоліків. Фінансові ризики мають інше походження і пов'язані з нестачею грошових сум для компенсації страхових подій, що виникли згідно укладеної страхової угоди. Звідси можна дійти висновку, що даний метод непридатний для оцінювання фінансових ризиків.

Регресійні моделі. Основою даного методу є пошук причинно-наслідкової залежності між індикаторами та рівнем ризику. В якості індикаторів, що спостерігаються, або пояснювальних змінних, можуть виступати наступні групи:

— змінні середовища, які являють собою кількісні міри, що описують бізнес операції підприємства;

— фактори ризику, які являють собою кількісні міри, що описують зразки реалізації ризику, що фіксувалися.

Математична модель подібного типу описується наступним чином:

$$y = Ax + b + \epsilon, \quad (1.7)$$

де y — обсяг втрат внаслідок настання фінансового ризику;

x — вектор змінних, що спостерігаються;

ϵ — випадкова величина, яка описує похибки або зовнішні збурення моделі;

A та b — параметри, які оцінюються в моделі, і описуються зв’язок між величиною y та спостереженнями x.

Для застосування даного підходу має бути достатньо даних з метою гарантії гарної точності оцінювання.

Методи нечіткої логіки. У ситуаціях, коли нема точних даних або їх недостатньо, даний підхід дозволяє використовувати експертні судження для оцінки ризиків. Нечітка логіка робить модель більш подібною до людини, коли вирішується питання щодо прийняття рішень та обґрунтування. Ризик неплатоспроможності страхової компанії та обсяги втрати можуть бути визначені за допомогою алгоритмів нечіткої логіки [32, 33].

Методи експертного оцінювання. Існують ситуації, що виходять за межі простих математичних задач під час дослідження складних систем, як фінансові. Використання людського мислення та його вміння вирішувати слабко визначені проблеми є базовим принципом даного підходу.

Алгоритм складається з таких кроків, які показано на рисунку 1.2.

Експерти обираються таким чином, щоб, в першу чергу, новообрані експерти повністю розуміли тонкощі бізнес-операцій чи предмет вибору і на додачу мали нульовий інтерес в результатах оцінювання. Звідси, групу експертів створюють з двох осіб, які є спеціалістами компанії та двох зовнішніх осіб. Завдяки аналізу професійної, наукової та іншої діяльності, кваліфікаційний рівень експертів в певній області знань застосовується для об’єктивної оцінки їх компетентності. Кожен експерт оцінює свої повноваження та своїх колег за апріорі визначеною шкалою як складова суб’єктивного підходу оцінки компетентності. Після обробки даних опитування визначається компетентність групи експертів разом з можливістю виявлення можливих помилок в оцінках.

Рисунок 1.2 — Схема функціонування методу експертного оцінювання

Виділяють такі методи обробки, які представлено нижче.

Метод надання переваг. Під час використання цього підходу, експерти впорядковують індикатори ризику за ознаками, де найменш характерний об’єкт

отримує одиничну оцінку. Під час отримання результатів, коефіцієнт відносно значущості елемента *j* обчислюється за формулою:

$$K_j = \frac{\sum_{i=1}^{n} K_{ij}}{\sum_{j=1}^{m} \sum_{i=1}^{n} ij}, \quad (1.8)$$

де K_{ij} — пріоритет характерності елемента *j* з точки зору експерта *i*;

n — число експертів;

m — число елементів, що аналізуються.

Показник, який має найбільше значення, стає тим показником ризику з точки зору експертів, що більш детально описує елемент.

Метод рангів. Використовуючи шкалу відносної релевантності, експерти оцінюють важливість кожного елемента у межах налаштованого діапазону, що зазвичай кратне 10. У противагу до попереднього підходу, цей дає можливість оцінити ступінь релевантності елементу *j* разом з його пріоритетністю.

Якщо у підприємстві нема необхідної інформації для виконання економічних та статистичних обчислень, то застосовуються експертні алгоритми, які використовуються з метою визначення ступеню фінансового ризику. Дані підходи ґрунтуються на опитуванні експертів з великим досвідом, результати яких оброблюються згодом математично.

Групові методи експертизи стали все більш популярними на практиці за останні роки. Дослідження продемонстрували, що методи групового генерування ідей дають більше пропозицій, аніж індивідуальні процедури експертизи від тих самих спеціалістів. Усі експертні підходи ґрунтуються на оцінці в балах окремих змінних ризику та визначення їх частини, незалежно від типу експертних процесів. Суб’єктивність результатних суджень є основним мінусом методу експертних оцінок.

Аналітичний метод. Цей підхід ґрунтується на загальновідомому принципу ринкової економіки. Його суть полягає в тому, що чим більший

ризик, тим більший дохід. Внаслідок цього, застосування будь-якого аналітичного алгоритму обмежує керування оцінкою приросту доходу проєкту та збільшення ризику проєкту або повної корисності. Існує безліч методів, що входять до даної групи, але зазвичай серед них виділяють метод аналізу абсолютних і відносних показників та метод аналізу чутливості.

Аналіз чутливості складається з таких етапів:

1) обирається основний індикатор, по відношенню до якого виконується аналіз чутливості;

2) виконується вибір факторів ризику, що виконують роль вхідних параметрів та мають здатність впливати на основний індикатор шляхом відхилення від очікуваного значення;

3) обчислюються більш точні значення основного показника на різних стадіях реалізації проєкту та для різних значень зафіксованих чинників ризику.

За рахунок потоків доходів та витрат, які задані даним чином, індикатори ефективності можуть бути обчислені в будь-який момент часу. Одночасно створюються діаграми, що демонструють зв’язок між змінними вхідними параметрами показниками, що виступають в якості результатів. Можливо виявити елементи, чиї модифікації мають найбільший вплив на остаточне значення ключових параметрів факторів ризику шляхом порівняння побудованих діаграм. Підприємець визначає недоліки проекту та створює стратегію для вирішення завдяки фіксації важливих значень змінних ризику. Наприклад, ціна проекту може бути знижена або маркетингова компанія може бути покращена, як за умови, що ціни на продукт виявляться найбільш важливим індикатором.

До головних недоліків даного підходу можна віднести його неповноту, що випливає з неможливості врахувати кожен сценарій, що може статися на момент реалізації. Також він не вказує ймовірність того, як будуть впроваджені інші ініціативи.

Метод аналогів. Інформація щодо того, як змінні ризику мали вплив на подібні проєкти, що були завершені, може бути корисною під час аналізу

ризику, що обмежує поточний. Це виконується за допомогою бази знань, де виводяться узагальнення та судження про виконання проекту. До недоліків даного підходу відносять описовий характер та реальність, що множина самих змінних ризику та їх факторів може варіюватися з плином часу.

Байєсівські мережі. Завдяки даному типу моделі можна описати причинно-наслідкову залежність між значущими чинниками ризику та змінами навколишнього середовища. У порівнянні з моделями регресії, мережі Байєса дозволять одночасно врахувати пряме відношення між змінними ризику та рівнем ризику так само як взаємозалежність між чинниками ризику. Також є додаткові умови для здійснення висновків з часткових даних при застосуванні даного підходу. З точки зору математики, Байєсівська мережа представляє собою орієнтований граф, де дуги виконують роль зафіксованих або очікуваних залежностей, а вершини виконують роль змінних ризику та зміни середовища. Також модель характеризує множину умовних розподілів випадкових величин, що представляють собою чинники навколишнього оточення та ризику.

Нехай є m випадкових величин типу $X_1,...,X_m$. Тоді повну ймовірність розподілу значень величин записується наступним чином:

$$P(x_1,...,x_m) = P(x_1)\prod_{k=2}^{m} P(x_k|(x_1,...,x_{k-1})). \quad (1.9)$$

До переваг Байєсівських мереж відносять їх здатність одночасного використовувати оцінки експертів для оцінки структури мережі за рахунок виявлення залежностей між змінними, вибору різних форм апріорних розподілів величин, та використання математичних методів для побудови висновків мережі. Як результат, модель дозволяє поєднувати знання експерта разом зі статистичною вибіркою даних.

Будь-який напрямок поширення інформації може бути використаний у висновку байєсівської мережі. При визначенні умовної ймовірності отримання значень для певних випадкових величин на основі відомих значень інших

змінних, дана модель використовується для побудови ймовірнісного висновку. Цю задачу можна описати математично як обчислення $P(y,x)$, де Y представляє множин змінних, що оцінюються, а X представляє множину змінних, що спостерігаються. Звідси, використовуючи наступний вираз, задача спрощується до визначення умовних ймовірностей:

$$P(y|x) = \frac{\sum_s P(y,x,x)}{\sum_{y,S} P(y,x,s)}, \tag{1.10}$$

де S — множина змінних, яка включає в собі X та Y.

Дані обчислення є витратними у плані часу, а побудова висновку є NP-повною. Множина стратегій для побудови висновку у Байєсівських мережах була розроблена з метою гарантії високої точності то полегшення механічних розрахунків.

Враховуючи наступні обставини, підходи до оцінювання ризиків можна умовно класифікувати за наступними групами, ґрунтуючись на всеосяжності інформації, до якої має доступ компанія:

— методи в умовах визначеності: дані щодо ризикової події є всеосяжною у вигляді звітів про прибутки та збитки, тощо;

— методи в умовах часткової невизначеності: дані щодо ризикової події доступні у формі частості цієї ситуації;

— методи в умовах повної невизначеності: відсутні знання щодо ризикової події, але є можливість частково усунути невизначеність за рахунок допомоги експертів та професіоналів.

Тому зниження початкового рівня знань щодо середовища діяльності робить вирішення задачі визначення рівня ризику більш складним, так як це зменшує можливість та надійність результатів.

Обчислювальні та аналітичні алгоритми або методи, що застосовуються для розрахунку показників ризику на основі даних звітності, використовуються

в умовах повної визначеності. У цих ситуаціях абсолютні, відносні та середні значення використовуються для надання індикаторних оцінок.

Абсолютні значення описують наслідки ризикових ситуацій:

— як частину балансу відношень, що демонструють результати фінансової та економічної діяльності;

— у вигляді матеріальних або вартісних виразів, якщо пов’язані втрати виступають суб’єктами даних одиниць виміру.

Ризик визначається у відносному представленні як потенційні втрати, що відносять до певної бази. Це полегшує врахування витрат ресурсів для конкретного типу підприємницької діяльності або очікуваний дохід в результаті діяльності. Вторинними щодо абсолютних індикаторів є відносні.

Поточні причини, чинники ризику та закономірності підприємницької активності відображаються в широких мірах середніх значень ризику. Водночас, за однією характеристикою, варіації результатів дій індивідуальних підприємців згладжуються та демонструють або загальну картину, або характеристику групи підприємців у певній сфері діяльності.

В умовах часткової невизначеності рекомендується застосовувати ймовірнісні та статистичні підходи оцінювання ризику, що розраховують ймовірнісні та статистичні індикатори оцінки ризику, оскільки ризик розглядається як ймовірнісна категорія. Ймовірнісні показники кількісно оцінюють ймовірність настання ризикової події та її наслідків. Зазвичай, обчислення ґрунтується на частості настання ризику, що потребує релевантних даних. Точкові або інтервальні оцінки використовуються для демонстрації пов’язаних результатів ризикової події.

Експертні оцінки, що характеризується високим ступенем суб’єктивності і надають допомогу у зменшенні рівня невизначеності та прийнятті свідомих рішень, можуть бути використанні у випадках повної невизначеності.

1.4 Види даних і невизначеностей

Статистичні дані, що були зібрані внаслідок досліджень або експериментів виконують роль об’єкту дослідження в прикладній статистиці. Набір елементів та їх характеристик, що їх задають, називають статистичними даними. Змінні, значення яких змінюються внаслідок досліджень, називають ознаками. Тип даних, що аналізуються, залежності між змінними та форми розподілу величин мають бути взяті до уваги під час вибору статистичного тесту. Для кількісної оцінки ознак процесу зазвичай застосовують такі чисельні шкали: номінальну, порядкову інтервальну або відносну.

У статистиці питання того як найкращим чином класифікувати дані все ще лишається актуальним. Але визнається те, що існують три різні класи даних: дискретні, неперервні та категорійні. Будь-які числові значення, які зазвичай виникають у певному порядку на числовій прямій відносять до неперервних. Злічена множина впорядкованих значень, що згодом групується за ранговою шкалою за ступенем фіксації, може бути віднесена до дискретних величин. Бінарні та категорійні значення можуть бути розділені серед категоріальних, що використовується для якісної категоризації та не придатні до сортування.

При моделювання та оцінюванні фінансових ризиків зазвичай присутні невизначеності. Їх можна розділи на наступні типи [34].

1. **Структурна невизначеність.** Виникає через нереальність встановлення причинно-наслідкових зв’язків та апроксимовані значення елементів структури. Для подолання зазвичай використовують методи експертного оцінювання, статистичні тести та теорію перевірки гіпотез.

2. **Статистична невизначеність при побудові моделей.** Пов’язана з похибками вимірювань, зовнішніми збуреннями, мультиколінеарністю, появою екстремальних значень та пропусками даних. З метою мінімізації використовують цифрові або оптимальні фільтри, уточнюють розподіл даних, застосовують метод головних компонент, теорію екстремальних значень та різноманітні методи заповнення пропусків даних.

3. **Параметрична невизначеність.** Можлива через некоректний вибір методу оцінювання параметрів або через малий об’єм даних. Для вирішення цього питання зазвичай забезпечують альтернативні методи оцінювання параметрів моделей або застосовують алгоритми розмноження даних.

4. **Ймовірнісна невизначеність.** Зазвичай присутня у ситуаціях наявності складних механізмів виникнення причинно-наслідкових відношень та відсутності детермінованості. З метою мінімізації даної невизначеності використовують статичні та динамічні Байєсівські мережі, моделі Маркова, ймовірнісні фільтри та умовні розподіли.

5. **Невизначеність амплітудного типу.** Пов’язана з наявністю змінних, що не піддаються вимірюванню. З метою зменшення ступеню впливу застосовують методи обробки нечіткої інформації.

6. **Інформаційна невизначеність.** Представляє собою неоднозначність процесів, явищ та інформації щодо системи, що досліджується сумісно з відсутністю знань щодо надійності інформації.

7. **Комбінаторна невизначеність.** Являє собою нездатність знати кожну можливу альтернативу. Усі інші форми невизначеності пов’язані з даним типом і зазвичай є його наслідком.

У подальших дослідженнях розглядалися наступні види невизначеностей, наведені в таблиці 1.1.

Таблиця 1.1 — Невизначеності, що розглядалися у дослідженнях

Тип невизначеності	*Пояснення*
Структурна невизначеність	Розглядалася проблема побудови структури регресійної моделі.
Статистична невизначеність при побудові моделей	Розглядався випадок екстремальних значень при побудові регресійних моделей.
Параметрична невизначеність	Розглядалася задача оцінювання параметрів регресійної моделі.
Ймовірнісна невизначеність	Розглядалась задача застосування мереж Байєса при оцінювання актуарних ризиків.

1.5 Числові характеристики вибірок

З метою вивчення та встановлення загальної динаміки статистичної вибірки використовується множина індикаторів, що носить назву описові статистики. До них відносять такі, представлені нижче.

1. **Вибіркове середнє.** Являє собою значення математичного сподівання що обчислюється за формулою:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{1.11}$$

2. **Мода.** Представляє собою значення вибірки, частота появи якої є найбільшою.

3. **Медіана.** Значення вибірки, яке у відсортованому вигляді більша або дорівнює одній половині і водночас менша або дорівнює іншій половині.

4. **Стандартне відхилення.** Величина, що дорівнює квадратному кореню з дисперсії і обчислюється наступним чином:

$$\sigma = \sqrt{Var(x)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{1.12}$$

де N — розмірність ряду;

$\bar{x}$ — вибіркове середнє;

$Var(x)$ — дисперсія ряду.

5. **Асиметрія.** Дана характеристика визначає ступінь симетричності розподілу або хвостів і має наступну форму для обчислювання:

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3. \tag{1.13}$$

Додатне значення цієї величини свідчить про “важкість” правого хвосту розподілу, від’ємне значення вказує на “важкість” лівого хвосту. У випадку нуля говорять про симетричність розподілу.

6. **Ексцес.** Дана характеристика визначає ступінь подібності розподілу, що розглядається, до нормального розподілу і обчислюється наступним чином:

$$K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4. \tag{1.14}$$

Значення $K = 3$ вказує на нормальний розподіл. При $K > 3$ кажуть, що розподіл є більш гострим за нормальний. $K < 3$ вказує про плоскість розподілу, що розглядається.

7. **Статистика Жака-Бера.** Даний показник характеризує ступінь близькості досліджуваного розподілу до нормального. За своєю структурою він являє собою різницю асиметрії та ексцесу вибірки, що досліджується і має наступну структуру:

$$JB = \frac{N-n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right), \tag{1.15}$$

де n — число параметрів, що оцінюється під час побудови моделі.

Статистика Жака-Бера має розподіл χ^2 з двома ступенями свободи при нульовій гіпотезі щодо нормальності. Ймовірність, що нульова гіпотеза є істинною виражається ймовірністю, що має зв’язок зі статистикою Жака-Бера. Низька ймовірність вказує, що нульова гіпотеза щодо нормальності повинна бути відхилена.

1.6 Аналіз існуючих рішень та технологій

Передові актуарні технології були швидко впроваджені та застосовані. У той же час нові технології продовжують одночасно розвиватися. У зв'язку зі швидкістю розвитку технологій та підходів, актуарії майже з великою ймовірністю щороку зіштовхуватимуться з новими інструментами, які вони мають розуміти та оцінити з метою покращення якості власної роботи, вдосконалювати ефективність наявних процедур та залишатися релевантними в бізнесі.

Згідно з інтерв'ю та результатами відкритого опитування, багато людей без сумніву думали про штучний інтелект (ШІ) [35]. Актуарії, котрі брали участь у дослідженнях, робили акцент на потенціал для допоміжних технологій як ШІ та машинне навчання для повного перетворення актуарної роботи. ШІ може використовуватися для вдосконалення прийняття рішень, автоматизації процедур та покращення моделювання ризиків. ШІ вже використовується декількома компаніями для оптимізації процедур андерайтингу, оброблення страхових позовів та виявлення шахрайства. З метою вдосконалення методів управління ризику та створення прогнозів, алгоритми машинного навчання можуть досліджувати дані з минулого. Деякі актуарії висловили стурбованість щодо ШІ та машинного навчання. Деякі вірять, що автоматизація може згодом замінити актуарні позиції вищого рівня разом з позиціями нижчого рівня у майбутньому. Також було хвилювання, що набір умінь, необхідних для етичного та успішного застосування моделей машинного навчання суттєво відрізнятиметься від того, що актуарії використовували в минулому.

Чат-боти в основному застосовуються в секторах підтримки сфери страхування для виконання більш стандартних завдань або врегулювання страхових випадків. Але протягом цих обмінів, чат-боти збирають важливі клієнтські дані, за умови їх збору та впорядкування, що дозволяє актуаріям та спеціалістам в області даних проводити аналіз, що ґрунтується на даних.

Неструктуровані дані використовують 44% опитаних актуаріїв, і ще 7% мають намір використовувати їх у майбутньому [36]. Неструктуровані дані як правило не містять організаційної структури та апріорі визначеної моделі даних. Записи розмов з клієнтами, інформація з веб-сайтів конкурентів, дані зі стрічок соціальних мереж або інших порівнянних джерел являють собою лише частину наведених прикладів. За рахунок використання natural language processing для аналізу даних взаємодії з кол-центром та оптичних зчитувань змісту для перетворення аналогових даних в електронний формат, деякі актуарії бачать можливість більш глибоко зрозуміти первинні причини припинення та відмови від страхових зобов'язань. Актуарії фіксували вплив важливих елементів досвіду на прибутковість до цього моменту, але додаткова інформація, отримана цими інструментами, має здатність пояснити результати, коли вони стануть очевидними.

Хоча ідея управління ризиками підприємства або Enterprise Risk Management (ERM) існує вже деякий час, зараз вона ще є більш важливою для успішного керування ризику будь-якої організації у світі, що змінюється швидкими темпами. Стійка структура ERM враховує усі ризики, що можуть негативно вплинути на цілі ефективності організації, як внутрішні, так і зовнішні, кількісні та якісні. Будь-яка організація можу досягти власних стратегічних цілей та розвинути сильну стійкість до ризику за допомогою такої структури, за умови її коректної розробки та управління. Використовуючи комплексні та перспективні техніки оцінювання, процес власної оцінки ризику та платоспроможності (ORSA) є прекрасною ілюстрацією системи інтегрованої системи ризик-менеджменту, що можна легко застосовувати до інших секторів як складову структури ERM [36].

Від підготовки даних та побудови моделей до оптимізації і створення узгоджених результатів, комплексні послуги компанії Deloitte охоплюють кожну фазу життєвого циклу моделювання [37]. Актуарії мають досвід у забезпечення повних послуг у розробці та обслуговування моделей великим міжнародним страховим компаніям. Вони починають шляхом взаємодії з

клієнтами для повного розуміння їх унікальних потреб та цілей. Поглиблені дискусії та семінари потрібні для збору всіх релевантних даних. Протягом багатьох років актуарії покращували свою систему моделювання для гарантії, що їх метод є коректним та ефективним. Вона пропонує всехооплюючі рекомендації щодо проектів моделювання з чітко визначеними процесами реалізації. Вони надають доступ клієнтам до найкращої актуарної експертизи за рахунок застосування навченої та різноманітної робочої сили. Їх команда складається з експертів з передових галузевих систем моделювання, що включає програмне забезпечення, як Prophet, AXIS, PathWise та MoSes. Для знаходження можливих вузьких місць та можливостей покращення, актуарії використовують Prophet Optimizer для огляду моделей. Їхня команда підсумовує результати, пропонує налаштування, що мають бути в пріоритеті та застосовує вдосконалення моделей в апріорі визначеному порядку. Вони створюють надійні моделі, що враховують нещодавні досягнення та практики в бізнесі за рахунок використання новітні актуарні методи та технології. Клієнти можуть повністю або частково автоматизувати процес генерування результатів за допомогою своєї команди.

Інструменти автоматизації актуарії компанії являють собою:

— **автоматизація налаштування моделей:** виконується оптимізація та автоматизація налаштування актуарних моделей з метою мінімізації ручної роботи та ризику помилок;

— **керування таблицями припущень:** виконується автоматизація процесів перетворення даних, створення та обслуговування таблиць припущення, що використовуються в актуарних платформах, що гарантує узгодженість та здатність до аудиту;

— **перетворення даних:** конвертуються дані про поліси та активи у формат, який потрібний для актуарних інструментів. також виконується групування та оптимізація даних трудомістких обчислень;

— **звітність за форматом solvency ii та irfs 17:** автоматизація складний актуарний процес, що включає в себе параметризацію моделі,

підготовка припущень, налаштування обчислень, аналіз чутливості та кінцеве знаходження маржі solvency ii.

Існують програмні продукти або рішення, які частково можуть покривати поставлену автору задачу, але вони не працюють в комплексі. Запропоноване автором рішення було розроблено та впроваджено вперше, тому немає результатів які були б отримані іншими авторами на сьогодні (маючи інформацію, яка розташована у відкритих джерелах і доступна до пошуку). Проміжні результати по кожному етапу представлені у розділах три та чотири.

Висновки до розділу 1

Виконано аналіз сучасного стану страхового ринку України. Результати демонструють, що попри кризові ситуації, як пандемія COVID-19 та повномасштабна війна, дана сфера продовжує функціонувати та адаптується до змін. Розглянуті поточні умови, з якими зіштовхується страховий ринок України, посилює необхідність у розробці нових підходів для реорганізації та покращення фінансового середовища.

Проаналізовано проблему оцінювання фінансових ризиків у сфері страхування. Розподіл ризику є першочерговою задачею галузі страхування, так як його підґрунтям є страховий ризик. Було продемонстровано різноманіття задач актуарного моделювання. Завдання визначення ймовірності розорення страхової компанії є одним із ключових в області страхової математики, і виконує роль основи для початкових страхових ідей оцінки фінансової надійності, де береться до уваги умова запобігання та відсутності банкрутства. Через це зростає потреба у застосуванні математичних та економічних засобів для моделювання та аналізу фінансових ризиків та у використанні актуарних моделей, що визначаються найбільш оптимальний обсяг страхових премій за встановленою ймовірністю розорення.

Продемонстровано, що значна кількість досліджень станом на зараз акцентована на задачах оцінювання страхових ризиків, в основному

фінансових, з застосуванням методів експертного оцінювання, методів нечіткої логіки, методів у сфері економіки, тощо. Але відсутні відомі дослідження математичних методів моделювання, що беруть до уваги одночасно пряму залежність рівня ризику від факторів ризик та залежність між факторами ризику, і застосування сучасного ймовірнісного підходу та інших методів інтелектуального аналізу даних з метою побудови логічних висновків за неповними даними. Це демонструє важливість застосування алгоритмів інтелектуального аналізу даних до складнощів даного класу.

Проведено дослідження щодо типів даних та розподілів, що використовуються в актуарному моделюванні, що відносять до узагальненого класу розподілів. Крім того, враховано базові ідеї та визначення описових характеристик даних, що можуть бути застосовані при дослідженні статистичних вибірок.

Зроблено огляд існуючих підходів оцінювання ризику, які розвиваються, та використовуються провідними страховими компаніями.

РОЗДІЛ 2 МЕТОДИ ОЦІНЮВАННЯ АКТУАРНИХ ПРОЦЕСІВ

Вибір методів оцінювання актуарних процесів грає важливу роль, оскільки це впливає на короткострокові так і на довгострокові наслідки компанії. У першу чергу обрані методи дозволяють захистити статки страхової компанії та уникнути фінансових втрат, що може поставити під загрозу майбутній стан фінансової організації. Крім того, це дозволяє актуаріям приймати більш інформативні рішення, які узгоджуються з реаліями фінансового середовища. Також правильно обрані методи дають можливість покращувати впевненість інвесторів та інших зацікавлених сторін.

При оцінюванні актуарних процесів використовувалися наступні методи.

2.1 Байєсівські мережі

2.1.1 Загальна методологія

Умова постійного зростання кількості даних робить задачу пошуку дійсно значущої інформації складною, особливо для аналітиків у спеціалізованих галузях навіть за умови, що усі об’єкти є легко та швидко доступними. Разом з оцінюванням величезних об’ємів даних та вибором коректних підходів та алгоритмів, ґрунтуючись на власному досвіді, експерт має застосовувати інформацію, зібрану з даних. У зв’язку з цим, при аналізі даних, спеціалісти вимушені мати справу з діапазоном невизначеностей, що включає в себе шуми або пропущені дані, збурення та різноманітні похибки. Перелічені обставини збільшують складність роботи в цілому та зазвичай веде до менш точних прогнозів або передбачень [5].

Одним з найбільш відомих підходів у моделюванні та прогнозуванні, що дає можливість мінімізувати невизначеність, є Байєсівська методологія. При використанні Байєсівських мереж для задач дослідження або процедур є можливість використовувати як експертні оцінки так і статистичні дані. Дана структура дозволяє представлення даних під час прийняття рішень в режимі

реального часу або в якості баз даних чи масивів даних. Також дані у даній моделі можуть бути дискретними або неперервними. Завдяки відображенню факторів процесу та високорівневій візуалізації формату причинно-наслідкової взаємодії виникає можливість повного розуміння взаємозв'язку. Також дана парадигма враховує різноманіття невизначеностей, що включає в себе статистичну, структурну та параметричну невизначеність. Неповні дані та приховані вузли можуть також бути застосовані при побудові моделей. Байєсівський метод повністю узгоджується людською поведінкою прийняття рішень під час дослідження різних процесів.

Байєссівська мережа являє собою пару $< G,B >$, де G — орієнтований ациклічний граф, що являє собою мережу відношень причинно-наслідкового формату. Остання змінна представляє множину параметрів, що визначає мережу. Складова B включає в себе $\Theta_{X^{(i)}|par(X^{(i)})} = P(X^{(i)}|par(X^{(i)})$ для всіх варіантів станів $x^{(i)} \in X^{(i)}$ та $par(X^{(i)}) \in Par(X^{(i)})$, де $Par(X^{(i)})$ являє собою батьківську множину $X^{(i)} \in G$. Кожна змінна $X^{(i)} \in G$ вважається за вершину. Якщо досліджується більше, ніж один граф, тоді застосовується вираз $Par^{G}(X^{(i)})$ для визначення батьків $X^{(i)}$ у графі G.

Повна спільна ймовірність Байєсівської мережі обчислюється за наступною формулою:

$$P_B\left(X^{(1)}...,X^{(M)}\right) = \prod_{k=1}^{M} P_B(X^{(k)}|Par(X^{(k)})). \quad (2.1)$$

Часто дану структуру відображають як граф з певними властивостями. Ребро з'єднує вершини або змінні, що утворюють Байєсівську мережу. Даний елемент, який з'єднує змінні, показує причинно-наслідкове відношення між ними. Неорієнтоване ребро свідчить про відсутність причинно-наслідкового зв'язку між вершинами. Граф класифікується як неорієнтований, якщо всі його ребра неорієнтовані. Наявність стрілки на ребрі свідчить про напрямок

залежності. Якщо кожне ребро у графі є орієнтованим, то граф класифікується як орієнтований.

Визначають такі види Байєсівських мереж, як представлено нижче [38].

1. Дискретні Байєсівські мережі. Являють собою тип мережі, де вершини виступають в ролі дискретних величини. Вони містять в собі такі характеристики:

— кожна вершина виконує функцію події, яка описується випадковою величиною, що містить декілька станів;

— усі вершини, що з'єднані зі своїми батьківськими вершинами, задаються за допомогою таблиці умовних ймовірностей або функцій умовних ймовірностей;

— ймовірності для вузлів, що не містять батьківських, вважаються безумовними.

2. Динамічні Байєсівські мережі. У цій мережі значення вершин змінюються з плином часу. При вирішенні задачі моделювання процесів залежних від часу, даний підхід вважається найбільш доречним. Застосування табличного представлення умовних ймовірностей робить пояснення нелінійних процесів простішим, що є однією з головних переваг.

3. Неперервні Байєсівські мережі. Події можуть мати будь-який стан в заданому діапазоні, а величини у цій мережі є неперервними. Звідси випадкова змінна матиме множину можливих станів або діапазон допустимих значень і нескінченну множину точок. У даному випадку, ймовірнісний розподіл визначається за рахунок функції щільності та ймовірнісної функції розподілу.

2.1.2 Метод Lauritzen-Spiegielharter

Коли структура побудована з використанням алгоритмів або експертів певної сфери, аналітик має в розпорядженні Байєсівську мережу з заздалегідь визначеною структурою. За рахунок призначення значень стану вузлам та розрахунку значень максимальної ймовірності для інших вузлів можна

пояснити кожен можливий сценарій. Ймовірнісний висновок — процес оцінювання стану вершини за апріорною ймовірністю значень інших вершин. Ця задача є складною та неоднозначною з обчислювальної точки зору. Кількість обчислень напряму залежить від вершин мережі та ребер, котрі з'єднують вершини один з одним. Причина виникнення невизначеності пов'язана з тим, що різні алгоритми ймовірнісного висновку дають різні вихідні значення. Дана ситуація є більш типовою у випадку великих мереж Байєса. Внаслідок цього, не існує універсального алгоритму, який здатний дати найкращі результати для всіх типів мереж [38].

LS метод, який ще називають підходом Lauritzen-Spigielharter, являє собою основу підходів до кластеризації. Структура Байєсівської мережі перетворюється у об'єднане дерево з метою застосування ймовірнісного висновку. Таблиці умовних ймовірностей вузлів моделі перераховуються послідовним чином з використанням алгоритму поширення повідомлення вгору та вниз по дереву [39].

Структура Байєсівської мережі для наочного застосування LS-методу наведено на рисунку 2.1.

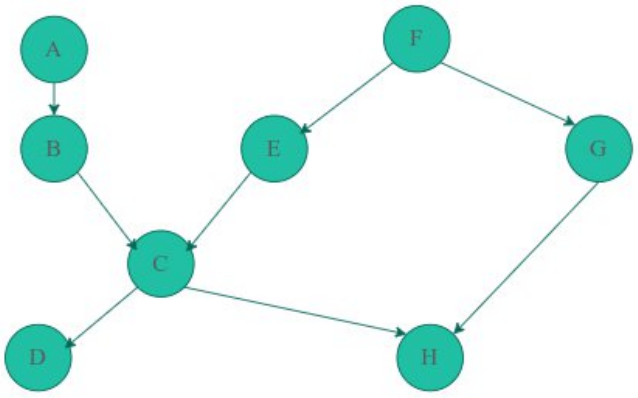

Рисунок 2.1 — Структура мережі Байєса для LS-методу

LS метод виконується у два етапи. Заповнення таблиць умовних ймовірностей вузлів моделі та побудова дерева об’єднань з базової структури мережі є складовою першого етапу. Він складається з п’яти кроків.

1. **Моралізація графу.** По черзі обирається вузол з батьками. Якщо батьківські вузли не з’єднані, застосовуються ребра з метою з’єднання взаємно розділених батьківських вузлів. Якщо батьківські вузли з’єднані між собою, то додається відношення сусідства. Дана модифікація часто носить назву “одруження” батьківських вузлів. Після цього усі орієнтовані ребра у даному графі мають бути замінені неорієнтованими ребрами, роблячи граф неорієнтованим. Приклад моралізації графу подано на рисунку 2.2.

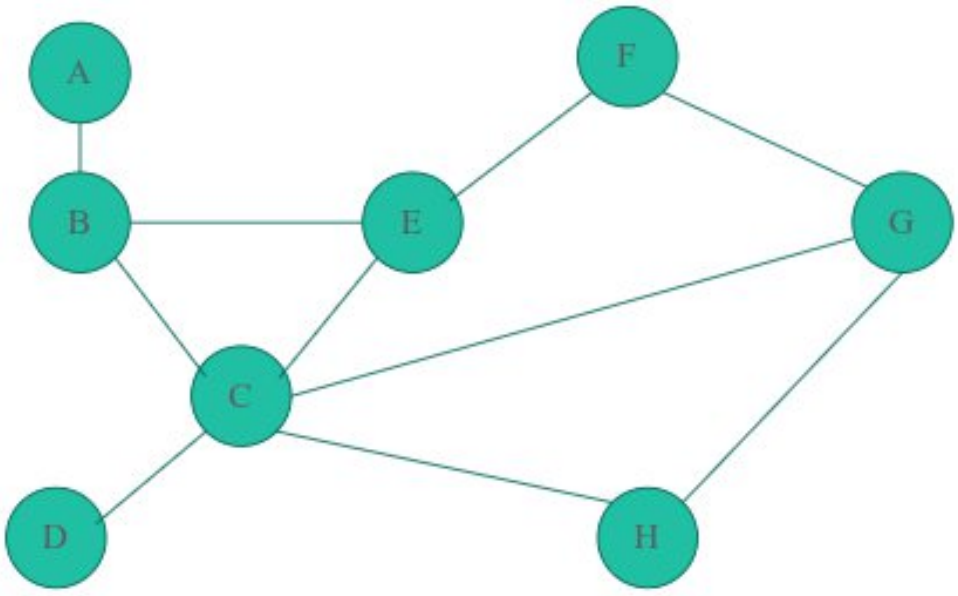

Рисунок 2.2 — Моралізований граф

2. **Триангуляція графу.** Проводиться перевірка, що кожен цикл морального графу містить не більше, ніж три вершини. В графічному сенсі це демонструється як ділення моралізованого графу на базові трикутники. Вершини Байєсівської мережі відвідуються одна за одною, поки не буде враховано кожен вузол. На цьому кроці необхідно:

— перевірити, якщо сусіди вузла, що розглядається, знаходяться в околі; якщо це так, вузол відносять до групи сімплиціальних, оскільки він формує кліку зі своїми сусідами і виключається з огляду разом зі своїми ребрами;

— після підтвердження факту, що мережа вузлів, які залишилися для розгляду, не є порожньою, відбувається перехід до наступного кроку метода; інакше, граф вважається триангульованим.

Послідовно аналізуються вершини мережі з метою знаходження вузла з максимальним числом сусідів. Цей вузол стає симпліціальним за рахунок додаванням додаткових ребер між власними несуміжними сусідами. Він вважається невідміченим і формує кліку зі своїми сусідами. Фундаментальний моралізований неорієнтований граф стає триангульованим за рахунок отримання нових ребер. Результат застосування процесу триангуляції наведено на рисунку 2.3.

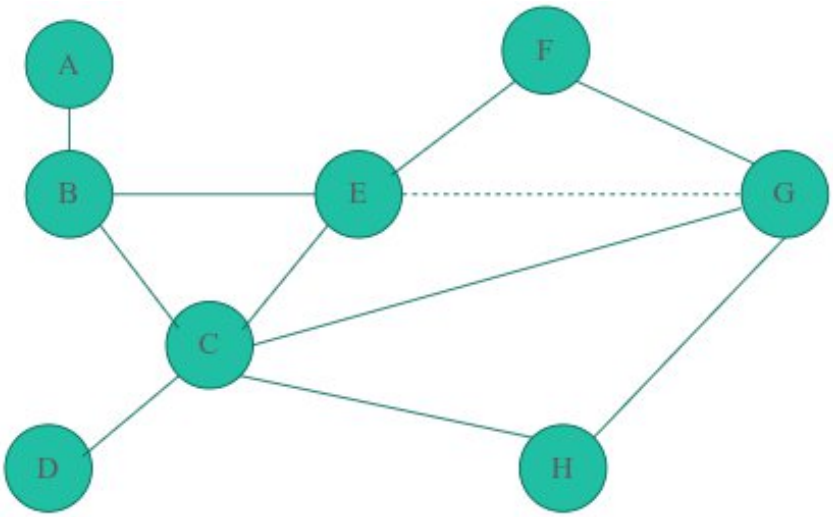

Рисунок 2.3 — Триангульований граф

3. **Виявлення клік.** Множина клік визначається за допомогою моралізованих графів або підграфів, потужність яких не перевищує трьох. Даний крок починається з визначення чи є граф, що розглядається,

підмножиною інших клік, які ще не були розглянуті. Якщо це так, вона буде відкинута. Якщо принаймні один вузол сумісно використовується цією клікою та іншими, то дуга, що містить сепаратор або перетин множин вузлів даних клік, додається між ними. Після цього здійснюється ранжирування множин клік за допомогою застосування впорядкованих множин вузлів. Перший вузол у кліці є тим, що має найвищий ранг у впорядкованій множині. У разі ситуації, коли декілька клік містять вузол, що займає другий ступінь у впорядкованій множині вузлів, проводиться перевірка його наявності. Внаслідок цього це застосовується до всіх множин клік. Набір клік для графу, що розглядається, наведено на рисунку 2.4.

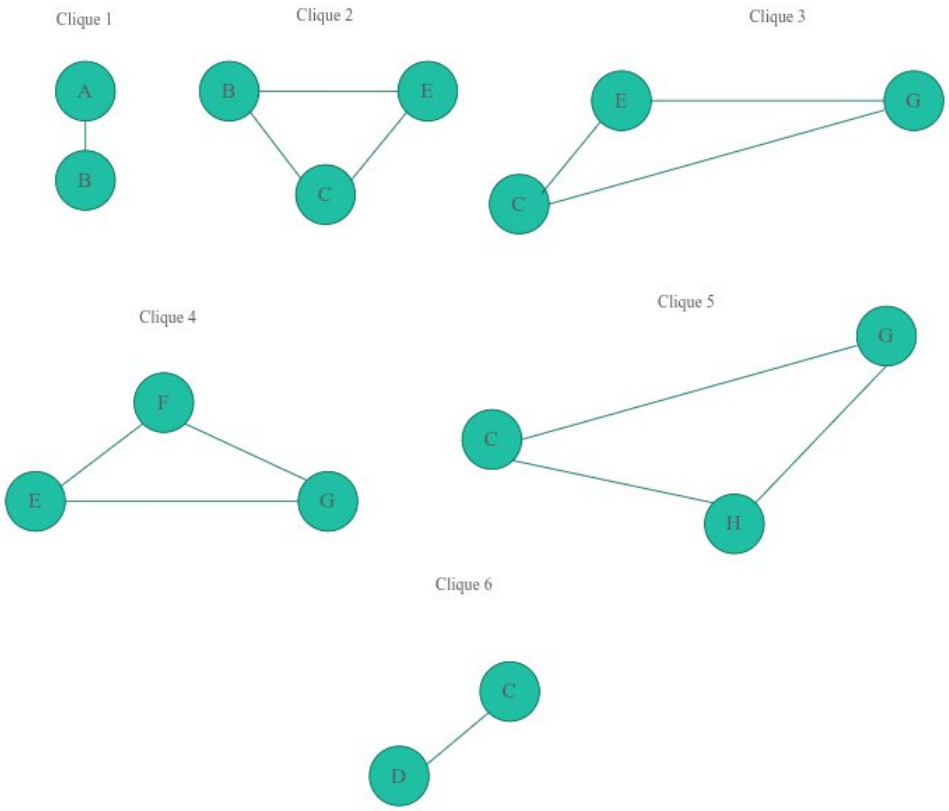

Рисунок 2.4 — Множина клік

4. **Побудова дерева об’єднань.** Обираються ребра з найбільшою кількістю сепараторів для черги об’єднання клік на даному кроці, у той час як

решта дуг видаляються. Кліка, що з'єднує усі кліки з найбільшим числом сепараторів, утворює дерево об'єднань. Для зручності кліка з найбільшою кількістю вершин в дереві об'єднань вибирається в якості кореня. Якщо існують декілька подібних вершин, обирається кліка з максимальною кількістю ребер.

Наприкінці першого етапу, дерево об'єднань заповнюється таблицями. Заповнення починається з листків дерева і продовжується по одній кліці за раз. В необробленій кліці враховуються невідмічені вершини:

— якщо є невідмічена вершина, що відсутня в інших необроблених кліках, таблиці слід мати структуру $P(vertex|\ other\ vertices\ of\ clique)$;

— якщо початковій структурі Байєсівськьої мережі не вистачає такої таблиці, використовується сумісна таблиця ймовірностей; кліки вважаються обробленими за умови, що всі вершини були відмічені;

— якщо мережа містить відмічені вершини і лише одну невідмічену вершину, таблиця структури $P(unmarked\ vertex)$ використовується з початкової структури Байєсівської мережі. Вузол вважається відміченим, коли кліка позначається як оброблена;

— застосовується таблиця сумісного ймовірнісного розподілу невідмічених вершин, і вони вважаються відміченими якщо є більше, ніж одна невідмічена вершина разом з відміченими;

— заповнюється таблиця структури $P(any\ vertex)$, і кліка вважається обробленою, якщо відмічені усі вершини необробленої кліки.

Утворене дерево об'єднань наведено на рисунку 2.5.

Алгоритм поширення, який виконує роль другого етапу, включає розрахунок ймовірнісного висновку з застосуванням дерева об'єднань. Для кожного спостереження обирається елемент з цією змінною. Усі інші відхилення від спостереження дорівнюють нулю. Процес руху вгору є першою фазою даного етапу. Коли змінні в таблиці, що відсутні в сепараторі додаються або маргіналізуються, результатом виступає повідомлення.

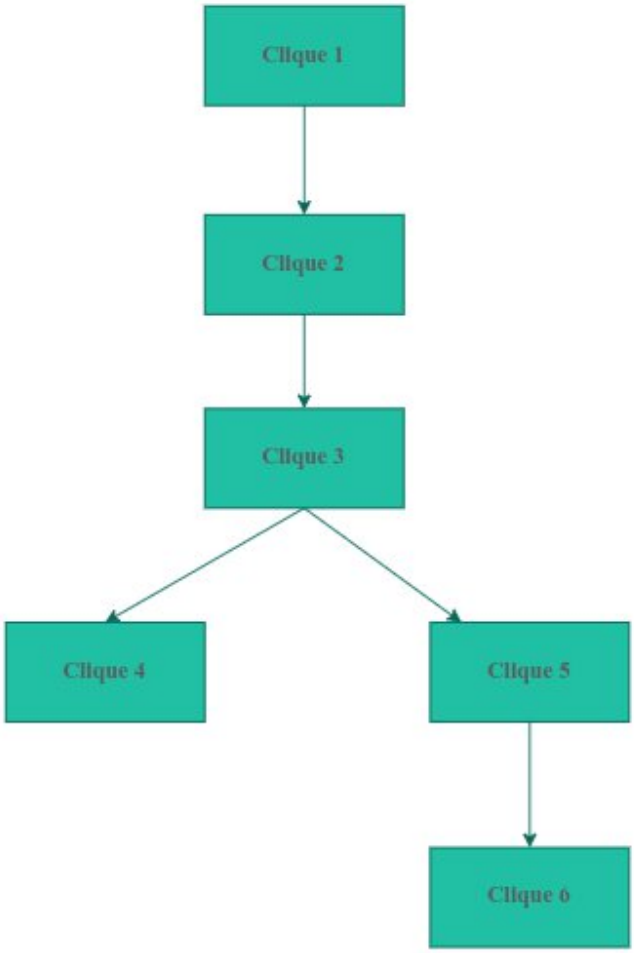

Рисунок 2.5 — Дерево об’єднань

Після надсилання повідомлення відправник ділить на нього свою таблицю умовних ймовірностей. За рахунок множення повідомлення на власну таблицю умовних ймовірностей, отримувач створює нову таблицю. Коли одержувач отримує нові повідомлення від власних дочірніх елементів, він ділить власну таблицю на отримане повідомлення і надсилає повідомлення власному батьківському елементу. Процес продовжується до тих пір, поки всі нащадки дерева об’єднань не надішлють повідомлення кореневому елементу.

Процес надсилання повідомлення вгору наведено на рисунку 2.6.

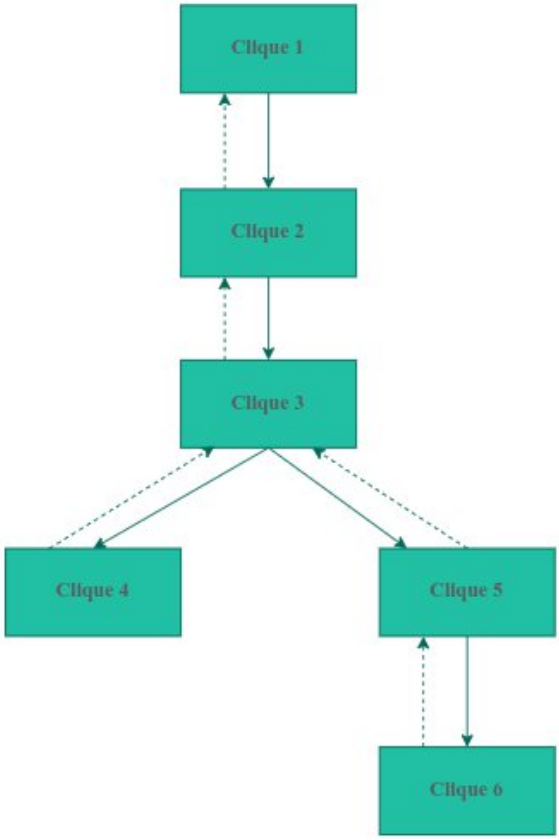

Рисунок 2.6 — Процедура надсилання повідомлення вгору

5. У міру просування процес руху вниз продовжується. Кореневий елемент відправляє повідомлення кожному нащадку. Він ділить свою таблицю умовних ймовірностей на повідомлення, отримане від нащадків, і маргіналізує таблицю за рахунок сепаратору перед передачею результату. До отримання повідомлення від власного батьківського елементу, нащадок множить його на власну таблицю умовних ймовірностей з метою генерування власної. Процес триває до тих пір, поки кожен листковий елемент не отримає власне повідомлення.

Процедура надсилання повідомлення вниз наведено на рисунку 2.7.

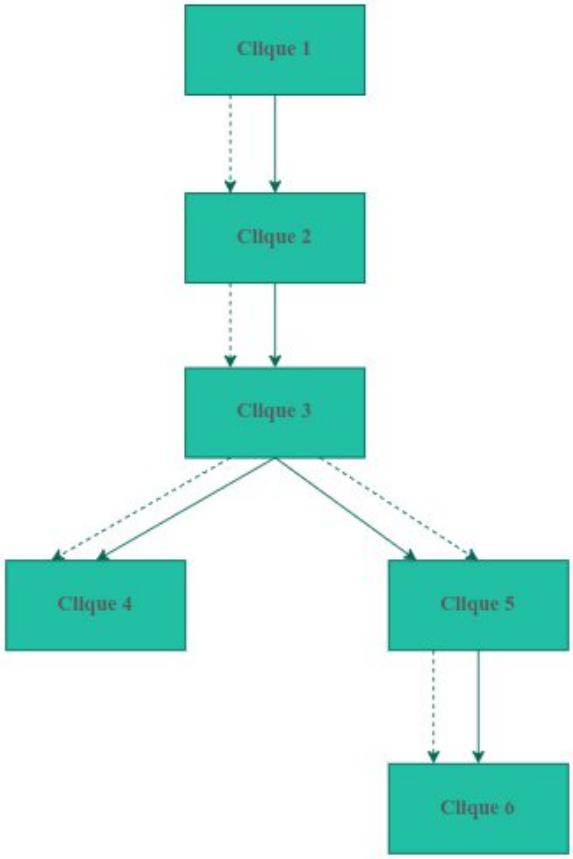

Рисунок 2.7 — Процедура надсилання повідомлення вгору

Результати будуть повторно обчислені в таблицях для кожної змінної, що має бути у подальшому нормалізовано за умови наявності спостережень.

2.1.3 Приклади застосування Байєсівських мереж у страхуванні

За рахунок унікальності побудови мережі Байєса та їх застосуванні при обчислюванні ймовірнісного висновку, дана модель використовується в різноманітних галузях діяльності. Фінансова область не є виключенням.

Сфери застосування мереж Байєса у страхуванні наведено на рисунку 2.8.

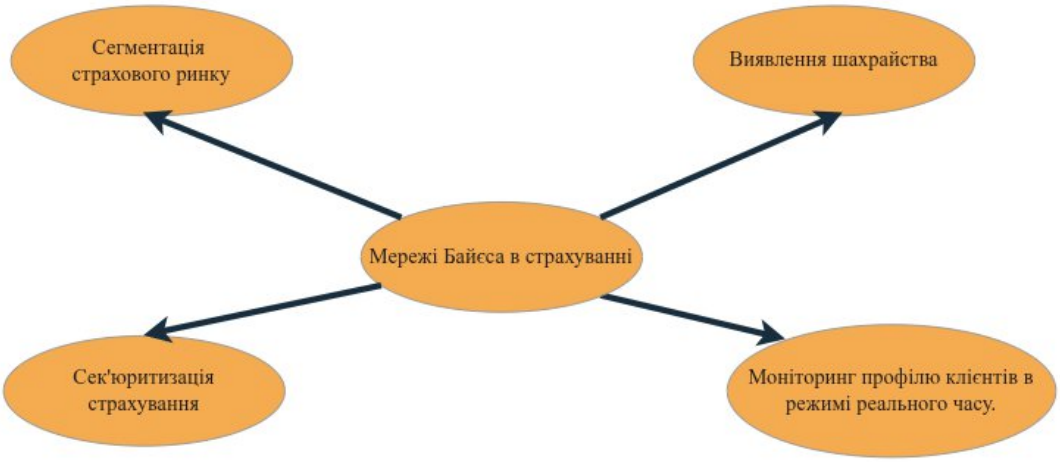

Рисунок 2.8 — Приклади застосування мереж Байєса в страхування

Завдяки впровадженню Байєсівської стратегії, спеціально створеної для страхування Pay-As-You-Drive (PAYD), дослідження [40] значно вдосконалило страховий сектор. За рахунок ациклічно-орієнтованого графу, Байєсівська мережа показала причинно-наслідкові зв’язки, у той час як наївна байєсівська модель покращила точність прогнозування розподілу ризиків. Страховики мають покращити свої тактики ціноутворення та забезпечити страхувальникам більш індивідуалізовані та справедливі рішення у страхуванні за допомогою включення результатів у поточну страхову практику. Їх дослідження мало декілька реальних застосувань. У першу чергу, покращена модель оцінювання ризику допомагає страховикам краще сегментувати власний ринок, залучаючи клієнтів більш високого рівня, яких відштовхують традиційні методи страхування. По друге, створення спрямованих планів профілактики та індивідуальних страхових полісів можуть мати напрямок на розуміння причинно-наслідкових зв’язків між величинами ризику. Модель PAYD може моніторити зміни в ситуації транспортування та керувальними звичками завдяки гнучкості до налаштування абсолютно нових даних. Насамкінець, корисним результатом, що може у майбутньому зменшити витрати та посилити цілісність страхового ринку є можливість вдосконаленого виявлення

шахрайства за допомогою виявлення справжньої взаємодії параметрів профілю страхувальників. Спрямовано-ациклічний граф, отриманий внаслідок аналізу авторами Байєсівської мережі, має потенціал для виявлення шахрайства у страхуванні, посилаючись на схожі дослідження, що включає в себе ідентифікацію та локалізацію діпфейків. Причинно-наслідкові зв'язки, виявлені ациклічно-орієнтованим графом, можуть бути застосовані для виявлення незвичних паттернів вимог, що відрізняються від звичайної керуючої поведінки та може свідчити про шахрайську діяльність.

Детальну структуру для подолання першочергової перешкоди, необхідної для сек'юритизації страхування ризику від нещасних випадків, тобто отримання точних і налаштованих остаточних прогнозів втрат для майбутніх років нещасних випадків, було надано Байєсівським процесом, описаним у [41]. Моделювання та реалізації в реальному світі, запропоновані авторами, продемонстрували, що байєсівська структура є водночас максимально прозорою та достатньо сильною для врахування складнощів, що лежать в основі цінних паперів, що мають відношення до страхування від нещасних випадків. Застосування байєсівських моделей наполягало, щоб вони перетворили кожне з припущень щодо прогнозування динаміки та розвитку втрат у математичну форму, що відкрита до детального вивчення та критики. Згодом автори застосували крос-валідацію з метою перевірки чи достатньо добре б функціонувала модель для практичного застосування. З точки зору авторів, дані алгоритми є необхідними для гарантії, що ціни на угоди щодо страхування від нещасних випадків встановлюються згідно методології, яка орієнтована на дані та вдосконалює відкритість та впевненість серед аналітиків та страховиків.

Хоча байєсівські мережі пропонували глибше розуміння ризиків клієнтів за рахунок моделювання ймовірнісних відношень між рисами та поведінкою клієнтів, підхід у [42] дозволив страховикам ефективно виявити важливі ознаки з великих немаркованих даних за допомогою самокерованого навчання. З метою того, щоб постійно випереджати постійно мінливий ринок, страхові

компанії мають оновлювати профілі своїх клієнтів в режимі реального часу та адаптивно. Байєсівські мережі та самокероване навчання роблять це можливим. Даний підхід гарантує довіру клієнтів та дотримання нормативних вимог, надаючи індивідуальні рішення, ґрунтуючись на даних. Значення точності та критерій AUC продемонстрували, що дана модель здатна підтримувати покращені здатності до прогнозу та розширити область ризик-менеджменту більш ефективно. Вона може призвести до більш персоналізованих страхових продуктів, а інтерпретованість, забезпечена ймовірнісною структурою байєсівської мережі, покращує впевненість при прийняття страхових рішень. Сфера страхування може значущим чином покращити довіру клієнтів та прозорість завдяки застосуванню самокерованого навчання та байєсівських мереж. Поки байєсівська мережа налаштовує ймовірнісні кореляції з метою надання можливості прозорого прийняття рішення, самокероване навчання використовує розріджені, невідмічені дані з метою вилучення корисної інформації, що дозволяє генерування індивідуальних правил. При їх сумісному застосуванні, вони покращують точність прогнозу, адаптуючись до зміни вимог клієнтів, і генерують результати, які легко розуміти, та сприяє раціональності та впевненості в процесі. Це б покращило обмеження точності, адаптивності, нестачі даних і одночасно б створило нові можливості для кращого обслуговування клієнтів та прийняття рішень. Безперечно, це представляє вдосконалення по відношенню до попередніх підходів.

2.2 Узагальнені лінійні моделі і оцінювання параметрів

2.2.1 Множина експоненційних розподілів

Традиційним методам утворення ставок не вистачає статистичної складності. Простий однофакторний та двофакторний аналіз зазвичай застосовується для аналізу досвіду страхових позовів за великою кількістю напрямків бізнесу. Хоча вони дають значущий прогрес, ітеративні алгоритми, що мають назву процедури мінімального зміщення та були розроблені

актуаріями у 60-і роки, все ще далекі від повної статистичної структури [43]. В дійсності, більшість популярних методів мінімального зміщення та класична лінійна модель є окремими випадками узагальнених лінійних моделей (УЛМ). Явні припущення щодо характеристики страхових даних та їх кореляції щодо прогнозних факторів можуть бути висунуті завдяки статистичній структурі УЛМ. Алгоритм розв'язання УЛМ є більш ефективним в технічному сенсі, ніж ітеративно-стандартизовані підходи, що є корисним на практиці, і водночас елегантним в теорії. Більш того, УЛМ пропонують статистичну діагностику, що допомагає перевірити гіпотези щодо моделі і обрати лише важливі змінні. В Європейському Союзі та на інших ринках УЛМ станом на зараз є загальновизнаними стандартними підходами у сфері для ціноутворення страхування малого бізнесу разом з приватними пасажирськими транспортними засобами та інших видів особистого страхування. Утворення ставок та андеррайтинг є головними областями страхового аналізу, де застосовуються УЛМ. Використання УЛМ для цільових маркетингових досліджень еволюціонувало внаслідок обставин, що обмежують здатність змінювати тарифи за бажанням.

Класична лінійна модель розглядається як окремий випадок УЛМ. Звідси доречно розглянути традиційні лінійні моделі з метою розуміння структури узагальнених лінійних моделей. Зображення зв'язку між спостережуваною змінною відгуку Y та коваріатами або предикторними змінними X є головною метою які лінійних так і УЛМ. Спостереження Y_i розглядаються в обох моделях як реалізації випадкової величини Y.

Лінійні моделі представляють Y як суму його математичного сподівання μ та випадкової змінної ε:

$$Y = \mu + \varepsilon. \tag{2.2}$$

Висуваються такі припущення:

1) очікуване значення Y, в ролі якого виступає μ, можна виразити у вигляді лінійної комбінації коваріат X;

2) випадкова змінна або похибка має Нормальний розподіл з нульовим математичним сподіванням та дисперсією σ^2.

Лінійну модель можна записати у вигляді наступної векторно-матричної форми:

$$\overline{Y} = E\left[\overline{Y}\right] + \overline{\varepsilon}, \tag{2.3}$$

$$E\left[\overline{Y}\right] = X\,\overline{\beta}\,. \tag{2.4}$$

Згідно дослідження [44] висуваються припущення, перелічені нижче.

1. **Випадковий елемент:** кожен елемент $\overline{Y}$ має Нормальний розподіл і є незалежним. Середнє значення кожного елемента може μ змінюватися, але вони всі повинні мати спільну дисперсію σ^2.

2. **Системний елемент:** лінійний предиктор отримується шляхом об'єднання p змінних:

$$\overline{\eta} = X\,\overline{\beta}\,. \tag{2.5}$$

3. **Функція зв'язку:** функція зв'язку визначає як випадковий та системний елемент, пов'язані один з одним. Тотожна функція та функція зв'язку в лінійній моделі є еквівалентними, тому:

$$E\left[\overline{Y}\right] = \overline{\mu} = \overline{\eta}. \tag{2.6}$$

Добре відомі методи лінійної алгебри можуть без проблем вирішити складні задачі, що виникають під час роботи з лінійними моделями. Водночас

стає зрозумілим, що доволі складно гарантувати необхідні припущення в застосуваннях [43]:

— для змінних відповіді, твердження щодо нормальності та постійності дисперсії є складним завданням; мета класичної лінійної регресії полягає у перетворенні даних таким чином, щоб ці припущення були істинними; так, $ln(Y)$ може відповідати припущенням, але Y може не відповідати цій вимозі;

— можливо обмежити значення змінної відповіді тільки на додатній діапазон; це обмеження порушується припущенням щодо нормальності;

— інтуїтивно зрозуміло, що дисперсія Y прямує до нуля, якщо змінна відповіді строго невід'ємна; інакше кажучи, дисперсія залежить від змінної;

— у більшості випадках застосувань адитивність ефектів, що враховано у припущенні щодо системного елементу та функції зв'язку, неможлива.

Лінійні моделі є окремим випадком серед багатьох моделей, що утворюють УЛМ. Нормальність, постійна дисперсія та адитивність ефектів, що виступають в ролі припущень обмежень лінійної моделі, відхиляються. Замість цього припускається, що змінна відповіді належить експоненційному сімейству розподілів. Також дисперсії дозволяється змінюватися разом з математичним сподіванням розподілу. Насамкінець, на трансформованій шкалі, припускається, що вплив змінних на змінну відповіді є адитивним.

Експоненційне сімейство розподілів можна представити формальним чином як двопараметричну множину як [43]:

$$f(y_i,\theta_i,\varphi) = exp\{\frac{y_i - b(\theta_i)}{a_i(\varphi)} + c(y_i, \varphi)\}, \quad (2.7)$$

де $a_i(\varphi)$, $b(\theta_i)$ та $c(y_i, \varphi)$ — функції, що визначені заздалегідь;

θ_i — параметр, що пов'язаний з середнім;

φ — параметр масштабування, що пов'язаний з дисперсією.

Доречно звернути увагу, що елемент експоненціального сімейства розподілів має наступні характеристики:

1) математичне сподівання та дисперсія повністю визначені;

2) дисперсія Y_i є функцією середньої.

Останню властивість можна описати наступною формулою:

$$Var(Y_i) = \frac{\varphi V(\mu_i)}{\omega_i}, \tag{2.8}$$

де $V(x)$ — заздалегідь визначена функція, що має назву функція дисперсії;

φ — параметр, що масштабує дисперсію;

ω_i — константа, що призначає вагу спостереженню i.

2.2.2 Основні складові узагальнених лінійних моделей

Нижче представлене є паралеллю до припущень лінійної моделі [44].

1. **Випадковий елемент:** Усі елементи $\overline{Y}$ є незалежними та належать експоненційному сімейству розподілів.

2. **Системний елемент:** Лінійний предиктор отримується завдяки комбінації p коваріат:

$$\bar{\eta} = X\,\overline{\beta}\,. \tag{2.9}$$

3. **Функція зв'язку:** диференційована та монотонна функція зв'язку, яка позначається g, використовується для встановлення зв'язку між випадковим та системним елементом таким чином, що

$$E\left[\overline{Y}\right] = \bar{\mu} = g^{-1}(\bar{\eta}). \tag{2.10}$$

2.2.3 Функція правдоподібності, функція зв'язку

Припускається, що розподіл кожної компоненти Y відповідає одному з правил сімейства експоненціальних розподілів. Загальна форма експоненціального розподілу дорівнює використанню канонічного параметру θ з метою виразу функції щільності розподілу за умови того, що φ відоме. У випадку нормального розподілу [44]:

$$f(y,\theta,\varphi) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} = exp\{\frac{\left(y\mu - \frac{\mu^2}{2}\right)}{\sigma^2} - \frac{1}{2}(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2))\}. \quad (2.11)$$

Звідси можна зафіксувати

$$\theta = \mu, \quad (2.12)$$

$$\varphi = \sigma^2, \quad (2.13)$$

$$a(\varphi) = \varphi, \quad (2.14)$$

$$b(\theta) = \frac{\theta^2}{2}, \quad (2.15)$$

$$c(y,\varphi) = -\frac{1}{2}(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)). \quad (2.16)$$

Запишемо наступний вираз $L(y,\theta,\varphi) = log\, f(y,\theta,\varphi)$ у вигляді логарифмічної функції правдоподібності як функцію, що залежить від y,θ,φ.

Математичне сподівання та дисперсія величини Y обчислюються наступним чином [44]:

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0, \quad (2.17)$$

$$E\left(\frac{\partial^2 L}{\partial \theta^2}\right) + E(\frac{\partial L}{\partial \theta})^2 = 0. \quad (2.18)$$

З формули (2.7) можна записати наступне:

$$L(y,\theta,\varphi) = \frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi). \tag{2.19}$$

Тоді отримуємо:

$$\frac{\partial L}{\partial \theta} = \frac{y - b'(\theta)}{a(\varphi)}, \tag{2.20}$$

$$\frac{\partial^2 L}{\partial \theta^2} = \frac{-b''(\theta)}{a(\varphi)}, \tag{2.21}$$

де $b'(\theta)$, $b''(\theta)$ — перша та друга похідна функції b від θ.

З виразу (2.17) та (2.20) випливає:

$$E\left(\frac{\partial L}{\partial \theta}\right) = \frac{\mu - b'(\theta)}{a(\varphi)} = 0, \tag{2.22}$$

звідси $E(Y) = \mu = b'(\theta)$.

Аналогічно для виразів (2.18), (2.20) та (2.21) дисперсії та другої часткової похідної функції правдоподібності по θ:

$$-\frac{b''(\theta)}{a(\varphi)} + \frac{Var(Y)}{a^2(\varphi)} = 0, \tag{2.23}$$

тобто дисперсія виводиться наступним чином:

$$Var(Y) = b''(\theta)a(\varphi). \tag{2.24}$$

Слід звернути увагу, що дисперсія Y складається з двох частин: дисперсійної функції, або $b''(\theta)$, що залежить від канонічного параметра і, в свою чергу, від середнього значення, та другої частини, що залежить більше від

φ, ніж від θ. $V(\mu)$ представляє функцію дисперсії, що розглядається як залежність від μ.

Як правило, $a(\varphi)$ записують наступним чином:

$$a(\varphi) = \frac{\varphi}{\omega}, \tag{2.25}$$

де φ — константа, яка також дорівнює σ^2 та має назву дисперсійного параметра;

ω — ваговий параметр, заданий апріорі та змінюється в залежності від досліду.

З іншої сторони, можна записати наступний вираз для нормальної моделі, де кожне випробування має математичне сподівання n:

$$a(\varphi) = \frac{\sigma^2}{n}, \tag{2.26}$$

де $\omega = n$.

При застосуванні традиційної лінійної регресії на практиці, дослідники можуть здійснити перетворення даних з метою забезпечення відповідності вимог адитивності ефектів, нормальності та постійності дисперсії змінної відповіді. З іншої сторони, УЛМ потребують лише функцію зв'язку, що гарантує остаточну вимогу адитивності. Припущення УЛМ щодо функції зв'язку вимагає, щоб деяка модифікація Y, виражена у формі $g(Y)$ була адитивною в контексті коваріат у противагу до припущення щодо функції зв'язку в лінійних моделях, що вимагає, щоб Y була адитивною в контексті коваріат [43].

На практиці зазвичай використовується обернене значення $g(x)$, оскільки краще досліджувати μ_i у формі функції лінійного предиктора:

$$\mu_i = g^{-1}(\eta_i). \tag{2.27}$$

Хоча в теорії можливо застосовувати окрему функцію зв'язку для кожного спостереження i, на практиці такий підхід використовується доволі рідко.

Функція зв'язку повинна бути монотонною, тобто строго зростаючою або строго спадною, та диференційованою. Математика, що використовується у розв'язанні УЛМ аналітичним способом спрощується за рахунок канонічної функції зв'язку, що пов'язана з кожною структурою похибки. Але застосування канонічних функцій зв'язку не є обов'язковим для розв'язання УЛМ з використанням сучасного комп'ютерного програмного забезпечення. Замість цього можна обрати будь-яку зручну комбінацію функції зв'язку та функції дисперсії.

2.2.4 Аналіз залишків та діагностика моделей

Змінну відповіді, що розглядається у моделі з нормальним розподілом, записують таким чином:

$$y = \hat{\mu} + (y - \hat{\mu}). \tag{2.28}$$

Інакше кажучи, даний вираз визначається як сума залишків та очікуваного значення. Елементи лінійного предиктора, функції зв'язку, функція дисперсії, адекватність прогнозованої моделі — усі вони проходять перевірку за допомогою залишків. З метою використання залишків при дослідженні інших розподілів, що можуть замінити нормальний розподіл, УЛМ вимагають більш детальної специфікації. Залишки є корисними з тієї ж причини, як і нормальний розподіл, що є зручним.

Розглянемо такі види залишків.

1. **Залишки Пірсона.** Цей тип має такий вигляд:

$$r_p = \frac{y-\mu}{\sqrt{V(\mu)}}. \tag{2.29}$$

Це початковий залишок без обробки, що обчислюється за допомогою потенційного відхилення Y. Термін походить з того факту, що залишок Пірсона для розподілу Пуассона дорівнює квадратному кореню χ^2-Пірсона, або якості моделі апроксимації даних таким чином, що

$$\sum r_p^2 = \chi^2. \tag{2.30}$$

2. **Залишки Енскомба.** Мінус залишків Пірсона полягає в тому, що вони показують асиметричні значення r_p для змінних, що не мають нормальний розподіл, що у свою чергу призводить до втрати деяких властивостей залишків, тобто характеристик, які зазвичай притаманні розподілу Гаусса. Енскомб запропонував опис залишків за допомогою функції $A(y)$ замість y, де $A(*)$ обирається для забезпечення найбільш досяжного нормального розподілу.

Веденберн показав, що функція $A(*)$ задається так, щоб функція правдоподібності з'явилася в УЛМ:

$$A(*) = \int \frac{d\mu}{\sqrt[3]{V(\mu)}}. \tag{2.31}$$

Для розподілу Пуассона вираз має такий вигляд:

$$\int \frac{d\mu}{\sqrt[3]{V(\mu)}} = \frac{3}{2}\mu^{\frac{2}{3}}, \tag{2.32}$$

тобто залишок потрапляє в діапазон: $\sqrt[3]{y^2} - \sqrt[3]{\mu^2}$

Хоча дане перетворення нормалізує функцію ймовірності, воно може не стабілізувати дисперсію через те, що процес включає ділення дисперсії $A(y)$ на її квадратний корінь, що виражається як $A'(y)\sqrt{V(\mu)}$. Залишки Енскомба для розподілу Пуассона визначаються наступним чином:

$$r_A = \frac{\frac{3}{2}(\sqrt[3]{y^2} - \sqrt[3]{\mu^2})}{\sqrt[6]{\mu}}, \tag{2.33}$$

Вілсон та Хільферт нормалізували змінну з розподілу χ^2 за допомогою застосування кубічного кореня.

3. **Залишки відхилення.** Збільшення значення d_i на одну одиницю в цій мірі призведе до того, що $\sum d_i = D$ буде відхиленням у вигляді міри невідповідності в УЛМ. Якщо задати r_D наступним чином:

$$r_D = sign(y - \mu)\sqrt{d_i}, \tag{2.34}$$

то можна отримати міру, що зростає як $y_i - \mu_i$ і виконується $\sum r_d^2 = D$.

УЛМ здатна забезпечувати значущу доповнену інформацію, показуючи достовірність тих параметричних оцінок, що є оцінками певного дійсного основного значення, сумісно з генерацією оцінок параметрів, що максимізують ймовірність.

Стандартні помилки. Невизначеність можливо оцінити за допомогою застосування статистичної теорії. Зокрема, стандартні помилки для кожної параметричної оцінки можливо визначити за рахунок багатовимірної нижньої границі Крамера-Рао, суть якої полягає в тому, що дисперсія оцінки параметра не менша мінус одиниці, поділеної на другу похідну логарифмічної правдоподібності [43]. Діагональний елемент матриці коваріації $-H^{-1}$, де H,

що має назву матриця Гессе, є матрицею других похідних логарифму функції правдоподібності, є тим, що мають на увазі під стандартними похибками. Стає інтуїтивно зрозуміло, що стандартні похибки є мірою того, як швидко логарифмічна правдоподібність відхиляється від максимального значення за умови зміни параметра.

Зазвичай припускається, що оцінки параметра розподілені за нормальним законом асимптотично, звідси теоретично можливо застосувати статистичний тест до кожної параметричної оцінки за допомогою порівняння з нульовим значенням, тобто визначивши, якщо вплив кожного рівня фактора суттєво відрізняється від базового рівня цього фактора. Квадрат оцінки параметра, поділений на його дисперсію порівнюється з розподілом χ^2 в тесті χ^2, що зазвичай застосовується для даного випадку. В дійсності, у цьому тесті порівнюються як параметр, так і базовий ступінь фактора. У зв'язку з довільною природою вибору базового рівня, цей тест може бути однозначно ефективним, якщо його застосовувати окремо. Трикутник χ^2-тестів, при порівнянні кожної пари параметричних оцінок, може бути потенційно побудований шляхом неперервної зміни базового рівня. Стає зрозумілим, що фактор не є значущим за умови, що одна з цих варіацій не є значущою.

Тести на відхилення. Міри відхилення можуть бути застосовані з метою оцінювання теоретичної значущості конкретної компоненти сумісно з стандартними помилками параметричних оцінок. Ступінь, до якої налаштовані значення відхиляються від даних, має назву відхилення.

Розглянемо функцію відхилення $d(Y_i,\mu_i)$, яка задається наступним чином [43]:

$$d(Y_i,\mu_i) = 2\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i-\zeta)}{V(\zeta)} d\zeta. \tag{2.35}$$

Припустимо, що застосовується УЛМ, що прогнозує μ_i, ґрунтуючись на самих спостереженнях Y_i. Коли функція дисперсії $V(x)$ мала, то значенню

$d(Y_i,\mu_i)$, тобто мірі відмінності між налаштованими та дійсними даними, надається більша вага. Інакше кажучи, будь-яка різниця між Y_i та μ_i підсилюється більшою мірою, якщо відомо, що Y_i походить з розподілу з низькою дисперсією.

Можна розглядати $d(Y_i,\mu_i)$ в якості узагальненого варіанту квадратичної похибки.

Загальна міра відхилення, яка носить назву загальне відхилення D знаходиться шляхом додавання функцій відхилення для кожного спостереження:

$$D = \sum_{i=1}^{n} 2\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i-\zeta)}{V(\zeta)} d\zeta. \tag{2.36}$$

Масштабоване відхилення $\check{D}$, яке можна розглядати як узагальнений варіант суми квадратів похибок, що бере до уваги форму розподілу, отримується за допомогою ділення на параметр масштабування φ:

$$\check{D} = \sum_{i=1}^{n} 2\frac{\omega_i}{\varphi} \int_{\mu_i}^{Y_i} \frac{(Y_i-\zeta)}{V(\zeta)} d\zeta. \tag{2.37}$$

Міри відхилення можуть бути застосовані для різних статистичних тестів. Коефіцієнт правдоподібності двох вкладених моделей, тобто моделей, в яких пояснювальні змінні в одній моделі виконують роль підмножини пояснювальних змін в іншій, є одним з найбільш корисних тестів. У противагу до тестів першого типу, що беруть до уваги значущість змінних, коли вони вводяться в модель послідовним чином лише з однією вільною змінною, що називається нульовою моделлю, ці тести носять назву тести третього типу.

Вибірка, що має розподіл χ^2 зі ступенями свободи, що дорівнюють різниці ступенів свободи між двома моделями може розглядатися як зміну масштабованого відхилення між двома вкладеними моделями, що представляє

відношення правдоподібностей. У цьому випадку, ступені свободи для моделі визначаються як число спостережень мінус число параметрів [43]:

$$\check{D}_1 - \check{D}_2 \sim \chi^2_{df_1 - df_2}. \tag{2.38}$$

Разом з нульовою гіпотезою, що додаткові параметри не є значущими, це дає можливість проводити тести з метою оцінювання важливості параметрів, що відрізняються між двома моделями. У своїй найбільш простій формі це оцінює чи додавання пояснювальної змінної до моделі покращує достатнім чином, тобто суттєво зменшує відхилення, враховуючи додаткові параметри, які вона вводить. Будь-який доданий фактор здатний покращити відповідність даним. Доволі важливе питання полягає в тому, коли покращення є суттєвим з урахуванням додаткової параметризації.

Масштабоване відхилення визначає результати тестів χ^2. Висувається припущення, що параметр масштабування відомий для певних розподілів, наприклад пуассонівський або біноміальний, і статистику можна обчислити з точністю. Параметр масштабу повинен бути розрахований для інших розподілів, так як він невідомий. Це, як правило, реалізовується за допомогою ділення на ступені свободи. Якщо оцінка параметра масштабу неточна, це може зменшити рівень надійності тесту.

Можна показати, що оцінка параметра масштабу розподілена за законом χ^2 після врахування ступенів свободи та дійсного параметра масштабу. Співвідношення розподілів χ^2 відомо як *F*-розподіл. Це в свою чергу веде до *F*-розподілу для відношення змін у відхилення до уточненої оцінки масштабу:

$$\frac{(D_1 - D_2)}{(df_1 - df_2)^{D_1}/_{df_2}} \sim F_{df_1 - df_2, df_2}. \tag{2.39}$$

Це свідчить про те, що *F*-тест можна застосовувати у випадках, коли

параметр масштабування невідомий, як у ситуації з гамма розподілом. Коли масштаб відомий, то використання даного тесту не має сенсу.

2.3 Порівняння альтернативних підходів до оцінювання актуарних ризиків

Лінгвістична змінна, що визначається виразом $< b,T,X,G,M >$, виконує роль представлення формату вектора даних з метою включення кількісних та якісних характеристик страхових клієнтів:

— b виконує роль назви лінгвістичної змінної;

— T представляє лінгвістичну фундаментальну множину термів, де кожен елемент являє собою нечітку множину, що задана на універсальній множині X;

— G функціонує як синтаксичне правило, яке дає можливість працювати з елементами термальної множини T та створювати нові терми; елемент носить назву граматика.

— функція належності нечітких термів, утворених синтаксичними правилами, утворюється завдяки семантичному правилу M.

Кожна функція належності лінгвістичної змінної має бути створена з метою встановлення правила [45]. Це передбачає визначення типу функції та її аргументів. Процес створення видає параметри функцій належності. Завдяки застосуванню моделі нечітких нейронних мереж, можливо автоматично створити функцію належності, задати базу знань та використати нечіткий логічний висновок. Такі кроки задають алгоритм:

1) задається множина вхідних змінних щодо страхового клієнта $\tilde{X} = \{X_1,...,X_M\}$;

2) задається набір вихідних змінних, елементами якого виступають потенційні групи ризику $D = \{D_1,...,D_N\}$;

3) створення фундаментальної терм-множини для кожного елементу суміжно з його функціями належності $A = \{a_1,...,a_N\}$;

4) визначення обмеженого числа нечітких правил, які мають сенс по відношенню змінних, що вони використовують;

5) застосування логічного висновку, або, інакше кажучи, виявлення нечітких підмножин для вихідних змінних кожного правила та перевірка дійсності передумов кожного правила;

6) побудова нечітких підмножин кожної вихідної змінної згідно правилам;

7) визначення точного значення для кожної вихідної лінгвістичної змінної.

Алгоритм Мамдані. Чотири фази реалізують логічний висновок в нечіткій нейронній мережі з логічним висновком Мамдані [46].

1. На початку методу вводиться нечіткість. Ступінь істинності задається для передумов правил: $A_1(x_0)$, $A_2(x_0)$, $B_1(y_0)$, $B_2(y_0)$.

2. Обчислення логічного висновку. Застосовуючи операцію пошуку мінімуму, визначається відсікання значення для передумов правил.

$$\alpha_1 = A_1(x_0) \cap B_1(y_0), \tag{2.40}$$

$$\alpha_2 = A_2(x_0) \cap B_2(y_0). \tag{2.41}$$

Далі визначаються граничні рівні функції належності:

$$C_1^{'} = (\alpha_1 \cap C_1(z)), \tag{2.42}$$

$$C_2^{'} = (\alpha_2 \cap C_2(z)). \tag{2.43}$$

3. Потім реалізовується етап композиції. Застосовуючи операцію максимуму, визначається об’єднання знайдених функцій належності з

відсіканням та використовується функція належності з метою отримання остаточної нечіткої підмножини для вихідної змінної.

$$\mu_{\Sigma} = C(z) = C_1'(z) \cup C_2'(z) = (\alpha_1 \cap C_1(z)) \cup (\alpha_2 \cap C_2(z)). \quad (2.44)$$

4. Застосовується обернена операція перетворення до чіткості. На даному етапі, як правило, використовуються центроїдний алгоритм.

Схема алгоритму Мамдані наведено на рисунку 2.9.

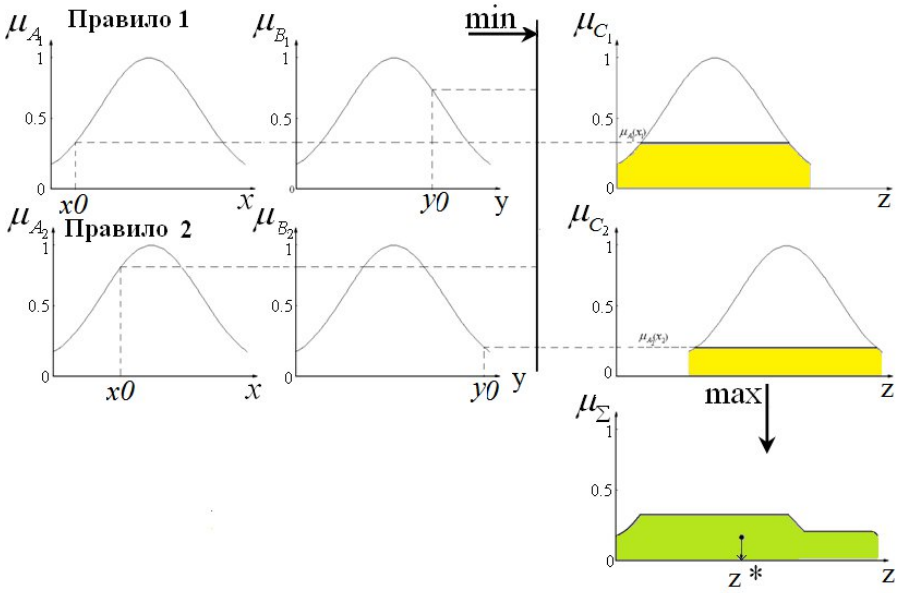

Рисунок 2.9 — Алгоритм Мамдані[1]

Ймовірність остаточної вартості страхових виплат в інтервальному вигляді є вихідною змінною, в той час як вхідні параметри включають в себе такі характеристики, що мають вплив, як вік, стать, режим роботи, кількість робочих днів, сімейний стан, тижнева оплата праці та кількість дітей. Для задачі оцінювання платоспроможності можливо застосувати наступну формулу: Кожна частина інформації щодо страхового клієнта забезпечується вектором типу $(x_1,...,x_n)$, де x_i виконує роль інформації, що була структурована у

[1] Джерело зображення: https://uk.wikipedia.org/wiki/Файл:Процедура_логічного_висновку_Мамдані.png

певному вигляді. Розмір страхових виплат має бути визначений в залежності від наданого вектору, що включає в собі оцінку ймовірності в заздалегідь визначеному інтервалі. Знаходження множини вхідних та вихідних змінних, створення базової множини термів з пов'язаними функціями належності для кожного терму, та побудова скінченної множини нечітких правил, що ґрунтуються на змінних, являють собою загальні етапи нечіткого логічного висновку. Критерій відсікання, що затверджується організацією, визначатиме точність моделі сумісно з числом помилок I-го і II-го типу. Завдяки встановленню критерію відсікання, можна виявляти не тільки частину клієнтів, яким надається найвищий пріоритет, але і ймовірність моменту, що страхова компанія збанкрутує. Моделі на основі нечіткої логіки можуть застосовуватися для оцінювання значущості клієнтів страхування та їх класифікації як більш або менш пріоритетних.

Алгоритм Сугено. Створення набору нечітких нейронних мереж стало результатом комбінування нейронних мереж з нечіткою логікою. Множина включає в себе мережу TSK (Takagi, Sugeno, Kang), яка ґрунтується на логічному висновку Сугено, мережу ANFIS (adaptive network-based fuzzy inference system), яка подібним чином ґрунтується на логічному висновку Сугено, тощо. Мережа TSK з висновком Сугено може бути використана для розв'язання задач оцінювання платоспроможності страхової компанії. Нечіткий висновок Сугено виглядає наступним чином. Множина правил в якості бази для логічного висновку задається наступним чином:

$$\Pi_1: if\ x \in A_1 AND\ y \in B_1, THEN\ z = a_1 x + b_1 y, \tag{2.45}$$

$$\Pi_2: if\ x \in A_2 AND\ y \in B_2, THEN\ z = a_2 x + b_2 y, \tag{2.46}$$

де x, y — вхідні величини;

z — результатна величина;

A_1, A_2, B_1, B_2, C_1, C_2 — функції належності, які визначені апріорі;

a_1, a_2, b_1, b_2 — зафіксовані числа [47].

Алгоритм складається з трьох кроків.

1. Застосування нечіткості, аналогічного до Мамдані.
2. Знаходження нечіткого висновку, що включає в себе обчислення:

$$\alpha_1 = A_1(x_o) \cap B_1(y_o), \tag{2.47}$$

$$\alpha_2 = A_2(x_o) \cap B_2(y_o), \tag{2.48}$$

а також окремі результати правил за допомогою виразів:

$$\dot{z}_1 = a_1 x_0 + b_1 y_0, \tag{2.49}$$

$$\dot{z}_2 = a_2 x_0 + b_2 y_0. \tag{2.50}$$

3. Розрахунок чіткого значення результатної величини:

$$z_0 = \frac{\alpha_1 \dot{z}_1 + \alpha_2 \dot{z}_2}{\alpha_1 + \alpha_2}. \tag{2.51}$$

Правила мережі продемонстровано наступним чином для надання більш зрозумілого та детального прикладу:

$$R_1: if\ x_1 \in A_1^{(1)}; \ldots; x_m \in A_m^{(1)};\ then\ y_1 = p_{10} + \sum_{j=1}^{M} p_{1j} x_j, \tag{2.52}$$

$$R_N: if\ x_1 \in A_1^{(N)}; \ldots; x_m \in A_m^{(N)};\ then\ y_N = p_{N0} + \sum_{j=1}^{M} p_{Nj} x_j. \tag{2.53}$$

У правилі, що використовує функцію належності, $A_i^{(k)}$ виконує роль значення лінгвістичної змінної:

$$\mu_A^k(x_i) = \frac{1}{1+\left(\frac{x_i - c_i^{(k)}}{\sigma_i^{(k)}}\right)^{2b_i^{(k)}}}, \tag{2.54}$$

Застосовуючи композицію отримуємо результат:

$$y(k) = \frac{\sum_{k=1}^{N} w_k y_k(x)}{\sum_{k=1}^{N} w_k}, \tag{2.55}$$

де $\mu_A^k(x)$ — ступінь, яка демонструє виконання правила

$$\mu_A^k(x) = \prod_{j=1}^{M} \frac{1}{1+\left(\frac{x_j - c_j^{(k)}}{\sigma_j^{(k)}}\right)^{2b_j^{(k)}}}. \tag{2.56}$$

Нечітка нейронна мережа Сугено містить п’ять шарів. Кожен з них виконує такі функції, які представлено нижче [46].

1. Кожна змінна x_i, $i = 1,...,M$ фазифікується окремо в першому шарі, що згодом обчислює значення функції належності $\mu_A^k(x_i)$ для кожного правила висновку k, ґрунтуючись на функції фазифікації. В якості прикладу можна застосовувати дзвіноподібну функцію. Параметри $c_j^{(k)}$, $\sigma_j^{(k)}$, $b_j^{(k)}$ даного параметричного шару мають бути налаштовані під час фази навчання мережі.

2. Для визначення ступеня належності w_k вектора X умови правила k, другий шар реалізує агрегацію окремих змінних x_i

$$w_k = \mu_A^k(x) = \prod_{j=1}^{M} \frac{1}{1+\left(\frac{x_j - c_j^{(k)}}{\sigma_j^{(k)}}\right)^{2b_j^{(k)}}}. \tag{2.57}$$

3. Функції $y_k(x)$ та w_k, що були розраховані в попередньому шарі, множаться, і значення $y_k(x) = p_{k0} + \sum_{j=1}^{M} p_{kj} x_j$ визначається в третьому шарі. Параметри p_{k0} та p_{kj} у даному шарі повинні бути налаштовані.

4. Зважена сума сигналів $y_k(x)$ та сума ваг $\sum_{k=1}^{N} w_k$ обчислюється в четвертому шарі.

5. Вагові коефіцієнти повинні бути нормалізовані в п’ятому шарі, результуючий сигнал повинен бути розрахований за виразом:

$$y(x) = \frac{\sum_{k=1}^{N} w_k y_k(x)}{\sum_{k=1}^{N} w_k}. \tag{2.58}$$

Результати кожного правила повертаються у входи мережі разом з вхідними величинами x_i в рекурентній нечіткій нейронній мережі TSK, що сумісно використовується для вдосконалення точності висновку. Для обчислення нових висновків, попереднє значення кожного виходу правила зберігається.

Схема роботи алгоритму Сугено наведено на рисунку 2.10.

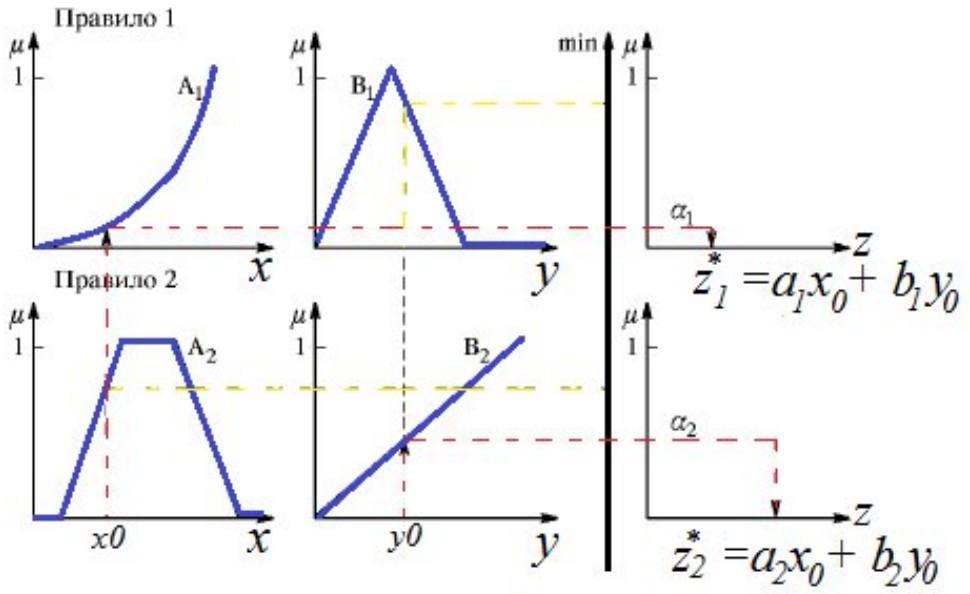

Рисунок 2.10 — Алгоритм Сугено[2]

2.4 Порівняльний аналіз методів і підходів

Порівняння характеристик УЛМ, мереж Байєса та нечіткої логіки подано в таблиці 2.1.

Таблиця 2.1 — Порівняння характеристик УЛМ, мереж Байєса та нечіткої логіки

Метод	*УЛМ*	*Мережі Байєса*	*Нечітка логіка*
Тип даних.	Дискретні та неперервні.	Дискретні та неперервні.	Дискретні та неперервні.
Здатність працювати з невизначеністю	Здатна.	Здатна.	Здатна.

[2] Джерело зображення: https://uk.wikipedia.org/wiki/Файл:Алгоритм_Сугено_1-порядку.png

Продовження таблиці 2.1.

Метод	*УЛМ*	*Мережі Байєса*	*Нечітка логіка*
Обчислювальна складність.	Висока.	Висока.	Середня.
Придатність для актуарних ризиків.	Придатна.	Придатна.	Придатна.
Рівень точності прогнозів.	RMSE, MAE, MAPE, Log-Likelihood, Байєсівський інформаційний критерій, інформаційний критерій Акайке,	Байєсівський інформаційний критерій, інформаційний критерій Акайке, ROC-AUC.	RMSE, MAE, R2, MAPE.
Гнучкість до динамічних прогнозів.	Здатний до адаптації.	Здатний до адаптації.	Здатний до адаптації.

На основі даних таблиці 2.1 можна зробити висновок, що розглянуті вище методи є тими підходами, які можна застосовувати при аналізі актуарних ризиків, так як вони працюють з будь-якими типами даних і придатні до роботи з невизначеностями.

Висновки до розділу 2

Вибір методів при аналізі страхових ризиків грає ключову роль, оскільки це впливає на короткострокові та і на довгострокові наслідки для будь-якої компанії.

В першу чергу було прийнято рішення використовувати мережі Байєса. Дана модель дає можливість представлення даних при прийнятті рішень в режимі реального часу. Мережа дозволяє використовувати експертні оцінки та і статистичні дані в дискретному або неперервному форматі. За рахунок представлення та високорівневій візуалізації формату причинно-наслідкової взаємодії можливе повне розуміння процесу. В якості алгоритму ймовірнісного висновку було вирішено використовувати LS-метод або метод Lauritzen-Spigielharter. Даний метод представляє собою основу кластеризації. Мережа Байєса при реалізації перетворюється в дерево об’єднань для подальшого обчислення ймовірнісного висновку.

Крім того в якості окремого методу було обрано підхід на основі узагальнених лінійних моделей. Їх особливість полягає в тому, що обмеження лінійних моделей, як нормальність, постійна дисперсія та адитивність ефектів, відхиляються. Замість цього висувається гіпотеза щодо належності змінних до сімейства експоненційних розподілів. Разом з цим для дисперсії скасовується обмеження щодо постійності разом з математичним сподіванням розподілу. Також припускається, що на трансформованій шкалі вплив змінних на цільову змінну є адитивним.

Разом з цим в якості альтернативи пропонується розглянути методи нечіткої логіки для оцінювання актуарних ризиків, а саме методи Мамдані і Сугено. Порівняльний аналіз усіх трьох підходів показав, що їх усі можна використовувати під час аналізу актуарних ризиків. Вони дають змогу працювати з даними в будь-якому форматі і вони здатні до роботи з різними видами невизначеностей.

РОЗДІЛ 3 СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ І МЕТОДИ ОЦІНЮВАННЯ ПАРАМЕТРІВ

3.1 Невизначеності даних та їх врахування

Невизначеність охоплює існування елементів, які не дають можливості мати детерміновані наслідки, але не є зрозумілим, як вплив цих факторів може мати на результати.

Різні типи невизначеностей, що охоплюють ситуаційні, політичні, суспільні, глобальні, тощо, характеризуються кількома змінними якостями, що визначають категорію невизначеностей. Для подолання перешкод, пов’язаних з прийняття рішень в ситуаціях невизначеностей, критично важливо визначити рівень аналізу та типи невизначеностей, що розглядаються [3].

Варто звернути увагу, що невизначеність часто обмежена нестачею глибокого розуміння щодо певного об’єкту. Дійсно, брак знань про стани об’єкта не є єдиним джерелом невизначеності. Також інколи можливо врахувати неоднозначність задачі та критеріїв, що застосовуються під час прийняття рішень.

У багатьох реальних сценаріях, ступінь складності прийняття рішень визначається числом альтернатив, а також різноманітністю та кількістю критеріїв, що застосовуються для оцінки цих варіантів. Реальні ризики та невизначеність мають бути враховані в кожній діяльності, що впливає на задачі, оскільки вони є складовою минулого, теперішнього і майбутнього розвитку процесу, що аналізується. Будь-яка економічна дія включає деякий ступінь ризику та невизначеності, але, не дивлячись на те, як керують ризиком, невизначеність неможливо повністю усунути. Взаємозалежності та непередбачені ситуації можуть виникати в будь-який момент часу. Такі неочікувані події можуть призвести до відхилення, що значно змінюють структуру даних [48].

Фактичні дані рідко існують без проблем. Численні проблеми, що могли не бути виявлені під час збору даних, як збій датчиків, проблеми передачі

даних, чи неправильне введення даних, можуть вести до пошкоджених даних. Датасети вважаються найважливішою компонентою при побудові моделей. Якщо дані не оброблені належним чином, то погіршується ефективність моделі. Це вирішує проблеми якості даних та враховує пов’язані з ним невизначеності. Тому підготовка даних є важливим кроком.

Набагато простіше виявляти та обробляти помилки, викиди і пропущені дані та враховувати невизначеності, коли дані попередньо обробляються. Це веде до більш приблизного аналізу та моделювання, коли дані є більш чіткими та точними. Разом з належною обробкою даних, моделі можуть працювати значно краще. Це гарантує, що вхідні дані моделі налаштовані таким чином, щоб алгоритм був здатний навчатися на них. Дані, які були оброблені коректно, гарантують наявність моделей, які легко розуміти. Набагато простіше розуміти зв’язок між ознаками, коли дані добре структуровані та легко зрозумілі. Належна обробка може допомогти видалити нерелевантні вхідні дані та шум, що в свою чергу може призвести до більш загальних моделей, які менш схильні до перенавчання. Підготовка даних дає можливість належного керування різними типами даних. Інформація, яка є організованою, напівструктурованою і неструктурованою часто виявляється в фактичних датасетах. Так як попередня обробка вирішує проблеми з якістю даних, покращує продуктивність моделі, обробляє діапазон форматів даних і враховує невизначеності даних, вона є важливим для отриманні ґрунтованої інформації та розробки точних моделей прогнозування [49].

Підготовка даних з увагою до масштабування ознак є важливим компонентом будь-якого дослідницького проєкту. Масштабування ознак є важливим для стандартизації діапазону незалежних величин в моделях та допомоги алгоритмів у більш швидкій збіжності та кращої роботи. Беручи до уваги різноманітну структуру датасету, масштабування його ознак гарантує, що жодний атрибут не має непропорційний вплив через свою величину.

За допомогою віднімання середнього значення та ділення на стандартне відхилення, метод стандартного масштабування стандартизує розподіл ознак,

забезпечуючи ознаки за стандартним відхиленням, що дорівнює одиниці, та центрованим навколо нуля. Хоча це може бути застосованим до будь-якого розподілу, цей алгоритм працює найкраще, коли ознаки мають нормальний розподіл. Гарантуючи, що кожна ознака знаходиться в одному масштабі, стандартне масштабування допомагає в стабілізації збіжності алгоритмів градієнтного спуску. Це можна показати наступним чином [50]:

$$z = \frac{x - \mu}{\sigma}. \tag{3.1}$$

Для нормалізації ознак до певного діапазону, зазвичай [0,1], використовується масштабування min-max. Для гарантії, що кожна ознака вносить однакову вагу у кінцеву проекцію, віднімається мінімальне значення і ділиться на діапазон або розмах. Даний підхід часто застосовується, коли розподіл не є Нормальним і є кращим за умов, коли параметри мають знаходитися у вказаному діапазоні [49].

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{3.2}$$

Для того щоб надати кожному зразку норму, що дорівнює одиниці, використовується нормалізація, що носить назву L2-масштабування. Даний підхід є кращим, коли є велика евклідова відстань між зразками. За допомогою нормалізації вектору ознак кожного зразка до евклідової одиничної довжини, даний алгоритм масштабування вдосконалює ефективність алгоритмів, які залежать від відстані меж зразками.

Як правило датасет має велику кількість полів, і їх значення не завжди виражаються як числа. При використанні моделей для прийняття рішень чи виконання схожих завдань, нечислові змінні роблять їх більш складними для автоматизації моделей. У даному випадку видалення змінних, що не містять числові значення може значно погіршити результат моделей. Як наслідок, під

час попередньої обробки, категоріальні дані мають бути перетворені у числовий формат. Потрібно розуміти наступні два типи категоріальних даних перед тим, як виконувати цей процес:

— ординальні дані: текстові дані, які вказують рівень;

— номінальні дані: текстові дані, які не представляють семантичний рівень.

Серед методів попередньої обробки категоріальних даних можна виділити представлені нижче [50].

1. **Label encoding:** Кожне окреме текстове значення в методі Label encoding конвертується у ціле значення згідно вказаній послідовності. Інакше кажучи, форма заміни в текстових даних в полі з n рядків виконується з використанням наступної формули, якщо множина унікальних значень у полі дорівнює $A = \{a_0, a_1, a_2, \ldots, a_{k-1}\}$

$$LEnc(b_i) = \begin{cases} 0, b_i = a_0; \\ 1, b_i = a_1; \\ 2, b_i = a_2; \\ \vdots \\ k-1, b_i = a_{k-1}; \end{cases} \tag{3.3}$$

де $i = \overline{1, \ldots, n}$.

Label encoding зазвичай застосовується у випадках, коли є лише дві унікальні значення в області категоріальних даних, не дивлячись на те, що вважається найбільш прямолінійним та найбільш зрозумілим методом кодування категорійних даних. Інакше кажучи, підхід Label encoding замінює текстові значення на чисельні без врахування будь-якого рівня, звідси цей метод найкраще працює лише коли $k = 2$. Індекс точності моделі зменшується внаслідок неправильної інтерпретації результатів моделі значень, встановлених Label encoding, які мають математичне значення. Доволі складно описати даний

підхід, як унікальний, так як в реальності, більшість категоріальних даних в полі даних мають більше, ніж два унікальні значення.

2. **One-hot encoding:** Використання Label encoding для номінальних даних не завжди може бути ефективним. Як було вказано раніше, даний підхід може вести до неточностей, так як він присвоює чисельні значення різної величини номінальним значенням одного рівня. One-hot encoding вважається ефективним методом у даній ситуації. Цей метод додає кожне окреме значення як поле для датасету. Інакше кажучи, k нових стовпців додаються до датасету, якщо одне з текстових полів містить k унікальних значень. Комірки цих стовпців $b_{i,j}$ містять 0 або 1. Наступний вираз визначає, коли приймати значення 0 або 1:

$$b_{i,j} = \begin{cases} 0, b_{i,j} \neq a_j; \\ 1, b_{i,j} = a_j, \end{cases} \tag{3.4}$$

де a_j — унікальні значення текстового поля;

$j = \overline{0,...,k-1}$;

$i = \overline{1,...,n}$.

Недоліком методу One-hot encoding є те, що він однозначно веде до отримання більшого набору даних, якщо поле, де виконується форма заміни даних, має багато унікальних значень. Звідси, підхід рекомендується використовувати, коли є невелика кількість унікальних значень.

3. **Dummy encoding:** Даний підхід та One-hot encoding схожі між собою. One-hot encoding створює k підходів для k унікальних значень, у той час як dummy encoding створює $k-1$ полів для k унікальних значень. Це є головною відмінністю між двома підходами. Інакше кажучи, перший рядок унікального значення в підході Dummy encoding містить лише 0 значень.

Вираз (3.5) генерує такі значення в комірках $b_{i,j}$ полів, що є результатом dummy encoding:

$$b_{i,j} = \begin{cases} 0, b_{i,j} \neq a_j \; or \; b_{i,j} = a_0 \\ 1, b_{i,j} = a_j \end{cases}, \tag{3.5}$$

де $j = \overline{0,...,k-1}$;

$i = \overline{1,...,n}$.

Враховуючи усі моменти, Dummy encoding в деякому сенсі кращий за підхід One-hot encoding.

Процес вибору підмножини відповідних характеристик з більшого набору оригінальних на основі апріорі визначених стандартів, роздільність класів або продуктивність класифікації відомий як вибір ознак. У додатках задач машинного навчання це грає доволі важливу роль. У задачах класифікації малих вибірок, коли число доступних навчальних зразків значно менше, ніж число ознак, вибір ознак має найвищий пріоритет. Типовою проблемою у більшості реальних застосунках є категоризація малої вибірки. У задачах класифікації малих вибірок, вибір ознак допомагає вдосконалити ефективність прогнозування за рахунок вирішення проблем перенавчання.

Рекурсивне виключення ознак (recursive feature elimination, RFE) є відносно новою технікою вибору ознак задача класифікації малих вибірок серед інших методик вибору ознак. Усуваючи найменш значущі ознаки, чиє видалення мало б найменший вплив на навчальні похибки, RFE націлений на вдосконалення узагальненої ефективності. Крім того, метод опорних векторів, як було продемонстровано їх непогане узагальнення навіть для класифікації малих вибірок, мають сильний зв’язок з RFE.

RFE намагається усунути слабкі та непотрібні ознаки, при цьому зберігаючи незалежні ознаки, не дивлячись на свій значний потенціал в задачах вибору ознак в малих вибірках. Як було відмічено, дві слабкі ознаки, які самі по

собі є малозначущими, можуть значно покращити ефективність комбінованим чином, а надлишкові ознаки можуть надати кращу роздільність класу. Тому, на продуктивність класифікації може негативно вплинути просте усунення слабких чи дубльованих ознак.

Дискримінантна функція в лінійно роздільній задачі являє собою [51]

$$g(\boldsymbol{x_i}) = \boldsymbol{w} * \boldsymbol{x_i} + b, \tag{3.6}$$

де $\boldsymbol{x_i}$ являє собою навчальні дані, $\boldsymbol{x_i} \in R^n$,

b виконує роль фактору зміщення,

$\boldsymbol{w}$ — ваговий вектор.

Непостійні змінні ξ_i, які кількісно оцінюють відхилення точки даних від ідеальної гіперплощини, можуть бути введені для лінійно-нероздільних задач. Метод опорних векторів задається шляхом мінімізації:

$$\Phi(\boldsymbol{w},\xi) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C, \tag{3.7}$$

$$y_i(\boldsymbol{w} * \boldsymbol{x_i} + b) \geq 1 - \xi_i x_i,\ \xi_i \geq 0, \tag{3.8}$$

де y_i відповідає цільовій змінній $y_i = \{1, -1\}$, $i = 1,...,m$.

Дана задача оптимізації розв’язується за допомогою двоїстої задачі:

$$g(\boldsymbol{x_i}) = \boldsymbol{w} * \boldsymbol{x_i} + b, \tag{3.9}$$

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j (x_i * x_j), \tag{3.10}$$

$$0 \leq \alpha_i \leq C,\ i = 1,...,m, \tag{3.11}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0, \tag{3.12}$$

де α_i — коефіцієнти Лагранжа.

Найменш важлива ознака, що пов’язана з найменшим ваговим значення в $\boldsymbol{w}$, зберігається RFE. Ваговий вектор для лінійного методу опорних векторів обчислюється з виразу:

$$W = \sum_{k \in SV} y_k \alpha_k \boldsymbol{x_k}, \tag{3.13}$$

Для нелінійного методу опорних векторів w_i обчислюється за наступною формулою:

$$w_i = \frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{K}(-i) \boldsymbol{\alpha}, \tag{3.14}$$

де $\boldsymbol{K}(-i)$ — обчислена матриця ядра, утворена після видалення ознаки i з вхідних даних $\boldsymbol{x}$.

На кожному етапі RFE рекурсивно видаляє ознаки і повторно ранжує ознаки, що залишились, за рахунок повторного навчання методу опорних векторів з використанням цих функцій. RFE просто видаляє ознаку, якщо вона слабка. Тим не менш, при комбінації з іншими ознаками, погана ознака може бути все ще значущою. Звідси, ефективність класифікації може бути погіршена шляхом видалення слабких або дубльованих ознак. Еквівалентне значення w_i використовується для оцінки ефективності класифікації після видалення передбачуваної слабкої ознаки для визначення її значущості. Навіть якщо ця характеристика має найменше значення w_i, вона зберігається, якщо після її видалення ефективність класифікації погіршується.

Коли змінні є номінальними, однією з найкращих статистик для оцінювання гіпотез є тест Хі-квадрат незалежності. На відміну від звичайних статистик, хі-квадрат або χ^2 може давати точну інформацію про те, які категорії пояснюють будь-які виявлені відмінності, разом зі значущістю будь-яких спостережуваних різниць. Тому ця статистика є однією з найбільш цінних

інструментів у дослідницькому наборі доступних аналітичних методів за рахунок кількості та якості інформації, що вона може запропонувати. Під припущеннями статистики розуміють умови, що мають бути виконані для коректного використання, як і буль-яка інша статистика. Крім того, в якості тесту значущості, χ^2 завжди повинен використовуватися сумісно з відповідним тестом сили.

Тест хі-квадрат, що має також назву як тест без розподілу, є непараметричною статистикою. Непараметричні тести повинні бути застосовані, коли до даних застосовується будь-яка з нижчеперелічених умов [52].

1. Усі змінні мають номінальні або порядкові рівні вимірювання.

2. Поки певні параметричні тести вимагають групи рівних або майже однакових розмірів, розміри вибірок груп, що досліджуються, не є рівними; для χ^2, групи мають бути рівними або нерівномірними в контексті розмірів.

3. Одне з наступних припущень параметричних тестів порушується вхідними даними, що були оцінені на рівні інтервалу або співвідношення:

— дослідник повинен застосовувати статистику, незалежну від розподілу, замість параметричної статистики, так як розподіл даних був значно асиметричним або загостреним, тобто параметричні тести припускають, що залежна змінна має грубо нормальний розподіл;

— гомоскедастичність або постійність дисперсії порушується даними;

— неперервні дані були розбиті на декілька категорій з різних причин; внаслідок цього дані більше не є інтервальними або відносними.

Формула для обчислення хі-квадрат має наступний вигляд:

$$\sum \chi^2_{i-j} = \frac{(O-E)^2}{E}, \tag{3.15}$$

де O — спостережувані значення, тобто дійсне число випадків в кожній комірці таблиці;

E — очікуване значення;

χ^2 — значення комірки хі-квадрат;

$\sum \chi^2$ — формульний вираз для сумування всіх значень комірок хі-квадрат.

Знаходження суми кожного рядка та стовпчика є першим кроком у визначенні χ^2. Для цих сум є граничні значення рядків і стовпців, які називаються маргіналами.

Знаходження очікуваних значень для кожної комірки є другим кроком.

Формула (3.16) використовується для розрахунку очікуваних значень хі-квадрат [52]:

$$E = \frac{M_R * M_C}{n}, \tag{3.16}$$

де E — очікуване значення комірки,

M_R — рядковий маргінал для комірки;

M_C — стовпчиковий маргінал для комірки;

n — повний розмір вибірки.

Статистика χ^2 отримується завдяки додаванню значень χ^2 комірок після їх розрахунків.

Ступені свободи в таблиці необхідні для того, щоб знайти з таблиці хі-квадрат рівень значущості статистики. Формула (3.17) використовується для визначення ступенів свободи для таблиці χ^2:

$$df = (number\ of\ rows\ -\ 1) * (number\ of\ columns\ -\ 1). \tag{3.17}$$

Ансамблевий алгоритм навчання, що має назву Випадковий ліс, використовується для розв’язання задач класифікації та регресії. Під час етапу навчання, метод створює діапазон дерев рішення та видає на виході клас, що

свідчить про позицію класів у випадку класифікації або середній прогноз кожного дерева у випадку регресії. Даний алгоритм машинного навчання функціонує за рахунок Bootstrap sampling. Навчальні дані використовуються для генерації випадкових вибірок за рахунок бутстрепу.

Розглядається датасет $D = \{(x_1,y_1),\ (x_2,y_2),\}..., (x_m,y_m)$ зі спостережень n, де бінарною цільовою змінною є y_i, а вхідною ознакою виступає x_i. Вихідний датасет D використовується для створення бутстреп-вибірки B, де $b = 1,2,...,B$. Дерево рішень T_b навчається за допомогою використання підмножини характеристик для кожної бутстреп-вибірки D_b. Найкраща ознака x_j обирається в кожному вузлі з випадкової підмножини ознак з метою поділу даних, де n виступає в якості кількості ознак в підмножині і $j \in \{1,2,...,n\}$. Для максимізації інформаційного приросту або зменшення неоднорідності Gini або ентропії обирається оптимальна ознака та точка поділу.

Для розбиття s, що ділить вузол t на два підвузли, t_L та t_R, інформаційний приріст має такий вигляд [54]:

$$\Delta G(s,t) = G(t) - (\frac{N_{t_L}}{N_t} G(t_L) + \frac{N_{t_R}}{N_t} G(t_R)); \tag{3.18}$$

$$G(t) = 1 - \sum_{i=1}^{n} p_i^2, \tag{3.19}$$

де p_i — ймовірність ознаки у навчальній вибірці;

N_t — розмірність навчальної вибірки;

N_{t_L} та N_{t_R} — розмірність розбитих підвибірок.

Як тільки дерево B було навчено, прогнози з усіх дерев об’єднуються для надання прогнозів щодо нового спостереження x. На основі більшості голосів, кожне дерево T_b прогнозує клас $\tilde{y}_b \in \{0,1\}$:

$$y = mode\{\tilde{y}_1,\tilde{y}_2,...,\tilde{y}_B\}. \tag{3.20}$$

Загальна неоднорідність втрат або неоднорідність Gini у зв'язку з розбиттям з використанням x_j по всіх деревах використовується для обчислення значущості ознаки x_j. Неоднорідність втрат ΔG_j^b для ознаки x_j розраховується для кожного дерева T_b:

$$\Delta G_j^b; = \sum_{t \in nodes\ using\ x_j} \Delta G(s_t, t). \tag{3.21}$$

Середнє зменшення неоднорідності по всіх деревах являє собою загальну значущість ознаки:

$$Importance\left(x_j\right) = \frac{\sum_{b=1}^{B} \Delta G_j^b}{B}. \tag{3.22}$$

Особливо корисним засобом для оцінювання та врахування статистичної невизначеності являє собою фільтр Калмана, що дає можливість оцінювати та прогнозувати стан динамічних процесів в режимі реального часу [7]. У цій ситуації, модель налаштовується на характеристики завжди доступних випадкових збурень та вимірювань шуму, використовуючи оцінки, обчислені в режимі реального часу, матриць коваріації обраних стохастичних процесів. Переваги доступних оптимальних процедур фільтрації включають можливість оцінювати невимірювані компоненти вектора стану, здатність обчислювати оптимальні оцінки змінних стану та їх прогнозів, можливість виконувати ефективне змішування даних, можливість явно включати статистичні властивості шумів вимірювань та збурень стану, і здатність оцінювати стани та деякі параметри моделі одночасно [3].

Фільтри Калмана використовуються для оцінки станів на основі лінійних чи нелінійних динамічних систем у форматі простору станів. Модель процесу визначає прогрес стану від часу $n - 1$ до часу n наступним чином [53]:

$$x(n) = Ax(n-1) + Bu(n-1) + w(n-1), \tag{3.23}$$

де $x(n)$ — вектор станів в момент часу $n > 0$;

A — матриця системної динаміки;

B — матриця коефіцієнтів керування;

$u(n)$ — вектор керувань в момент часу $n > 0$;

$w(n)$ — вектор шумів в момент часу $n > 0$, що має нормальний розподіл з вектором середніх, що дорівнюють нульовому значенню, та матрицю коваріацій Q.

Зв'язок між станом та вимірюванням на поточному кроці n описується моделлю вимірювання сумісно з моделлю процесу наступним чином:

$$z(n) = Hx(n) + v(n), \tag{3.24}$$

де $z(n)$ — вектор вимірювань вихідних змінних в момент часу $n > 0$;

H — матриця коефіцієнтів спостережень;

$v(n)$ — вектор випадкових значень шуму вимірювання в момент часу $n > 0$, що має нормальний розподіл з вектору середніх, що дорівнюють нульовому значенню, та матрицю коваріацій R.

Функція фільтру Калмана полягає у наданні оптимальної оцінки $x(n)$ в момент часу n з врахуванням початкової оцінки $x(0)$, послідовність вимірювань $z(1),z(2),...,z(n)$ та специфіку моделі систем, визначної A, B, H, Q, R.

Хоча матриці коваріацій призначені для фіксації статистики шумів, у більшості реальних застосувань справжня статистика шумів або невідома, або не є гаусівською. Звідси Q та R зазвичай застосовуються як параметри налаштування, які користувач може задавати для отримання бажаної ефективності фільтра.

Даний метод також використовує матрицю коваріацій помилок оцінки

вектора стану, яка пов’язана з оцінкою стану. Вона позначається як $P(n)$.

Нижче представлено алгоритм фільтрації Калмана.

1. Задаються початкові умови $x(0)$ для вектора стану і матриці коваріації помилок оцінки $P(0)$. Присвоюють значення для збурень стану матриць коваріації Q та помилкам вимірювань R.

2. Застосовується такий вираз для знаходження оптимального коефіцієнту фільтра:

$$K(n) = \hat{P}(n-1)H^T\left[HP(n-1)H^T + R\right]^{-1}. \tag{3.25}$$

3. Використовуються нові спостереження для отримання поточної оцінки вектора стану:

$$\tilde{x}(n) = A\tilde{x}(n-1)(n)\left[z(n) - H\widetilde{A}x(n-1)\right] + K. \tag{3.26}$$

4. Обчислюється апостеріорна матриця коваріації помилок для оновлених оцінок:

$$P(n) = [I - K(n)H]\,\hat{P}(n), \tag{3.27}$$

де I — одинична матриця.

5. Проводиться пошук апріорної коваріаційної матриці помилок оцінки для наступної оцінки вектора стану:

$$\hat{P}(n+1) = AP(n)A^T \tag{3.28}$$

і відбувається перехід до кроку 2, тобто обчислень рівнянь фільтра.

Схема роботи фільтру Калмана наведено на рисунку 3.1 [55].

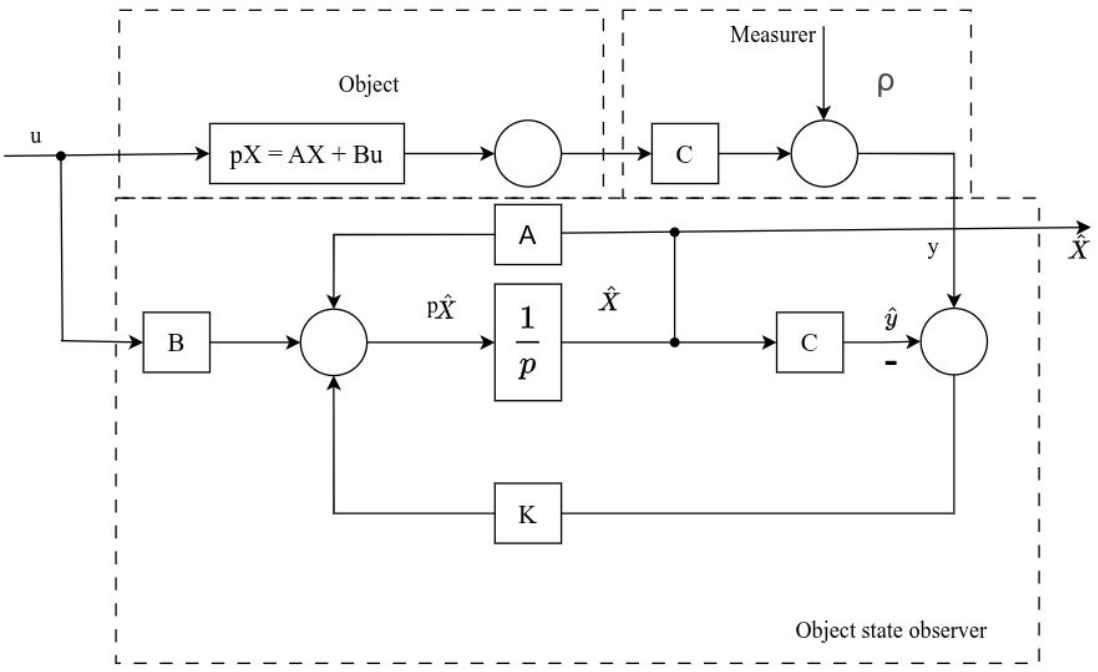

Рисунок 3.1 — Схема функціонування фільтру Калмана

Модель (3.29) і (3.30) виконує функцію спостерігача, представлена таким чином:

$$p\hat{X} = A\hat{X} + Bu, \tag{3.29}$$

$$\hat{y} = C\hat{X}, \tag{3.30}$$

де $\hat{X}$ — оцінка вектора стану об’єкта, який є зазвичай невідомим;

A — матриця стану;

B — вхідна матриця;

C — вихідна матриця;

u — вектор керуючих сигналів;

$\hat{y}$ — вектор вихідного сигналу, що вимірюється.

Спостерігач записується у вигляді виразу:

$$p\hat{X} = A\hat{X} + Bu + K(y - \hat{y}). \quad (3.31)$$

3.1.1 Проблема шуму та відсутності даних

Велика кількість факторів, що включає в себе джерело даних, розмірність вибірки та метод збору інформації, має великий вплив на якість даних у реальних додатках. Не дивлячись на значні зусилля щодо запобігання помилок, збір даних схильний до них. Існування шуму в даних може збільшити складність для побудови моделей машинного навчання, навіть за умови, якщо багато підходів машинного навчання включає в себе вбудовані стратегії обробки шуму, як процес відсікання для дерев рішень. Збільшена складність навчених моделей, більша тривалість часу обробки та потенційне зниження прогнозних властивостей для абсолютно нових даних представляють лише частину проблем. Ці моделі можуть зіштовхуватися з проблемами надійності та безпеки у разі застосування при стресових ситуаціях [56].

Існує декілька визначень шуму в літературі, присвяченій машинному навчанню та статистиці. Більшість з них приходять до висновку, що шумні дати можуть бути перешкодою для навчання, так як вони містять похибки. Хоча якщо викиди являють собою екстремальні або нетипові, але точні приклади, вони часто розглядаються як шумні дані під час дослідження. Автори [57] виділили дві категорії шуму в наборах даних навчання з учителем: шум у цільовому атрибуті і шум у прогнозних атрибутах. Останнє трапляється в класових мітках і може виникати в результаті помилок, суб’єктивного маркування або навіть неякісної процедури розмітки. Автор іншого дослідження [58] стверджував, що, наприклад, некалібровані метрики можуть призвести до систематичних неточностей у прогнозуванні якостей.

Процес автоматичного виявлення шуму є складним за структурою. Перед тим, як елементи, позначені як шумні, будуть оброблені у подальшому, в ідеальний ситуації має бути валідаційний етап, на якому підтверджується їх дійсність. Так як видалення шумних даних є доволі типовим методом, важливим є належним чином відокремлення цих даних від безпечних. Як тільки безпечні дані містять характеристики, що роблять внесок до інформації, що вимагається для побудови потрібної моделі, вони мають зберігатися в безпеці.

Оцінювання того, чи є певний зразок шумним чи ні в реальних додатках зазвичай вимагає думки експерта певної галузі, який не завжди доступний. Разом з цим, найм експерта до того, що робить етап попередньої обробки більш дорогим та витратним в контексті часу. Складність полегшується, коли використовуються штучні набори даних, або штучно згенерований шум додається до набору даних в режимі керування. Методичне включення шуму робить легшим перевірку методів виявлення шуму та аналізу того, наскільки шум вливає на навчання [56].

Так як більшість статистичних методів не були створені для обробки неповних даних, пропущені значення викликають серйозні перешкоди для дослідників певної галузі під час статистичного аналізу та інтерпретації даних, особливо коли припущення щодо ігнорування пропущених даних не справджується. Зміщенні оцінки як правило є наслідком неналежних підходів, що використовуються для корекції пропущених даних. Значення, що не спостерігаються, які також мають назву пропущених даних, виникають, коли дослідники не здатні виміряти реальні дані, які вони мають намір виміряти, що залишає неповну інформацію. В опитуваннях та випробуваннях, пропущенні дані зазвичай виникають, коли щось відбувається не так з невідомих причин. Існує багато різних причин, чому дані могли бути відсутні; деякі з них знаходяться під частковим керування дослідника, інші — ні. Учасники зазвичай здатні ігнорувати будь-які питання щодо чутливих тем чи навіть призупинити дослідження в будь-який момент часу, що значним чином впливає на появу

пропущених значень. Це пояснюється тим, що етичним чином вимагається, щоб участь в будь-яких дослідженнях була добровільною [59].

Дані можуть бути відсутні з різних причин, що включає ненавмисний пропуск елементу, неправильна реєстрація даних, нерозуміння проблеми або втрата інтересу всієї задачі, відмова відповідати на конкретне питання з певних причин, переїзд до іншого міста, що робить складним продовження поточного дослідження, або дані були пропущені навмисно. Розуміння того як і чому значення були пропущені, грає важливу роль, незалежно від підходу, що був застосований для врахування пропусків. Як описано в [60], відсутність даних можуть бути пов'язані з однією або комбінацією наступних причин: випадкові процеси; процеси, що вимірюються; процеси, що неможливо виміряти. У [61] відмічено, що визначення точного процесу, який надає пропущені значення допомагає у виборі найкращого підходу до обробки відсутніх даних.

Коли ймовірність відсутності значення даних не залежить від оцінки одиниці будь-якої змінної, не дивлячись на те, чи спостерігаються інші змінні чи ні, дані називаються повністю випадково відсутні (missing completely at random, MCAR), тобто

$$P(missing\ data \mid data\ are\ observed, data\ are\ not\ observed) = P(missing\ data).$$

Так як відсутність зазвичай виникає внаслідок інших факторів в датасеті, припущення MCAR є строгим та ірраціональним, і воно рідко виконується в сценаріях реального світу.

Коли значення відсутні навмисно, можна висунути припущення, що вони відсутні повністю випадковим чином. Відсутність вважається повністю випадковою, коли ймовірність $P(X[i] \in [a,b]\ not\ observed)$ в спостереженні i незалежна від значення x_i змінної X або від значення будь-якої іншої змінної в датасеті, тобто [59]:

$$\forall i \in O: P(X[i] \in [a,b]\ is\ not\ observed) = P(X[i] = unavailable\ value) \qquad (3.32)$$

де $O = \{1,2,...,n\}$ та $P(X[i] = \ unavailable\ value)$ — ймовірність спостереження $X[i]$ відсутнє, незалежно від значення x_i.

Коли ймовірність відсутності даних залежить від інформації, що доступна, ніж від неспостережуваних даних, то говорять, що дані відсутні випадково (missing at random, MAR). Схожим чином, коли відсутні дані випадково розподілені в межах однієї або кількох підвибірок, ніж випадково розподілені по всіх спостереженнях, це відомо як MAR. Допоміжні змінні в одному з випадків пояснюють причину того, чому значення відсутні або прогнозують оцінку для відсутніх значень, що можна застосовувати для підтвердження припущення MAR, яке є менш строгим, ніж припущення MCAR. Декілька допоміжних змінних мають здатність прогнозувати відсутні значення та надавати пояснення, чому вони були відсутні.

Ймовірність того, що x_i не спостерігається в інтервалі $[a,b]$, виражається формулою $P(missing\ data \mid data\ are\ observed, data\ are\ not\ observed) =$
$= P(missing\ data \mid data\ are\ observed)$, маючи на увазі, що $P\left(x_i \in [a,b]\ is\ not\ observed\right)$ не залежить від значення x_i змінної X, як було проконтрольовано значення іншої змінної Y в датасеті, тобто [58]

$$\forall i \in \{i \in O : Y[i] = c\}, \qquad (3.33)$$
$$P(X[i] \in [a,b]\ not\ observed) = P(X[i] = \ unavailable\ value),$$

де $\{i \in O : Y[i] = c\}$ — підмножина спостережень $Y[i]$ величини Y, де Y дорівнює деякому значенню c, що виступає в ролі константи.

Так як незміщені значення параметрів можуть бути виведені завдяки множинного заповнення або прямого методу максимальної правдоподібності без включення явної моделі, що пояснює, чому дані відсутні, вимогу MAR часто називають ігнорованою відсутністю.

Серед механізмів відсутності, припущення щодо невипадкової відсутності (missing not at random, MNAR) є найбільш проблематичним, так як

пропущені значення не розподілені випадково по всіх спостереженнях, а також цей розподіл не є випадковим в межах будь-якої підмножини, що може бути вилучена з датасету.

Оскільки відсутність залежить від самого відсутнього значення, неможливо визначити кількісним чином або спростити ймовірність відсутності, або $P(missing\ data\ |data\ are\ observed, data\ are\ not\ observed)$, за змінними в моделі.

Значення x_i, що не спостерігається у змінної X, визначає ймовірність $P(X[i] \in [a,b]\ not\ observed)$ [58]:

$$\forall\ i \in O, \tag{3.34}$$

$$P(X[i] \in [a,b]\ is\ not\ obsreved) = \frac{P(X[i] = unavailable\ value \cap X[i] \in [a,b])}{P(X[i] \in [a,b])},$$

де $P(X[i] \in [a,b])$ — ймовірність того, що змінна X потрапляє в інтервал $[a,b]$ в спостереженні i, незалежно від того відсутнє дане спостереження чи ні;

$P(X[i] = unavailable\ value \cap X[i] \in [a,b])$ — ймовірність того, що спостереження $X[i]$ відсутнє, знаходячись в інтервалі $[a,b]$.

Важко або навіть неможливо оцінити ймовірність пропущених значень, коли відсутнє спостереження залежить від речей або подій, що не були виміряні дослідником. Неігнорована відсутність — термін, який використовується для опису механізму невипадкової відсутності.

3.1.2 Методи врахування нелінійності

У подальших дослідженнях в якості методу врахування нелінійності цільової змінної використовувався тест Фішера.

Його статистика має такий вираз [34]:

$$\tilde{F} = \frac{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} n_i(\bar{y}_i - \tilde{y}_{ij})^2}{k-2}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n-k}}, \quad (3.35)$$

де k — кількість підвибірок;

n_i — розмірність підвибірки;

$\bar{y}_i$ — вибіркова середнє підвибірки;

$\tilde{y}_{ij}$ — значення, отримане в результаті регресійної моделі;

n — загальна розмірність вибірки.

Тобто дану статистику можна інтерпретувати наступним чином:

$$\tilde{F} = \frac{Deviation\ of\ mean\ value\ from\ regression\ model}{Deviation\ of\ y(k)\ from\ the\ mean\ values\ of\ the\ subsamples}. \quad (3.36)$$

Гіпотеза про лінійність відхиляється за умови виконання того, що розрахована статистика $\tilde{F}$, яка має ступені свободи $v_1 = k - 2$ та $v_2 = n - k$, не менша рівня значущості.

3.1.3 Методи врахування нестаціонарностей

Запропоновані УЛМ враховують ситуацію гетероскедастичності, так як страховий ринок, як і будь-який фінансовий ринок є волатильним.

В економетриці часові ряди зазвичай фінансового характеру з умовними дисперсіями, що залежать від попередніх значень, значень дисперсії та інших величин аналізуються за допомогою моделі авторегресійного умовної гетероскедастичності або АРУГ (ARCH). З моментами високої волатильності, що можуть тривати певний час, за якими слідують періоди низької волатильності та середньою волатильністю, як зазвичай є довгостроковою та

безумовною, яка може вважатися як достатньо стабільною, модель АРУГ будується для розв'язання задачі кластеризації в фінансових ринках [62].

Дана модель має таку структуру:

$$\tilde{\varepsilon}^2(k) = \alpha_0 + \sum_{i=1}^{q} \alpha_i \tilde{\varepsilon}^2(k-i) + v(k), \quad (3.37)$$

де $\tilde{\varepsilon}^2(k)$ — квадрат оцінки похибки моделі;

α_0 — параметр запізнення;

α_i, $i = 1,...,q$ — параметри моделі;

$v(k)$ — білий шум з нульовим середнім.

За рахунок регресійних моделей знаходяться похибки моделей.

Головним недоліком моделей АРУГ є те, щоб умовна дисперсія завжди має бути додатною, то параметри моделі повинні бути невід'ємними.

Згідно моделі узагальненої авторегресійної умовної гетероскедастичності або УАРУГ (GARCH), зміни в показниках та апріорних оцінках дисперсії матимуть вплив на поточні зміни в дисперсії [63]. Умовна дисперсія описується як процес авторегресії з ковзним середнім, що є по собі розширенням моделі УАРУГ. Вираз (3.38) використовується для характеристики похибок:

$$\varepsilon(k) = v(k)\sqrt{h(k)}, \quad (3.38)$$

де $\sigma_v^2 = 1$;

$h(k)$ — умовна дисперсія, яка записується таким чином:

$$h(k) = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2(k-i) + \sum_{i=1}^{p} \beta_i h(k-i), \quad (3.39)$$

де p — число минулих оцінок, що мають вагу на теперішнє значення;

β_i — параметри, що показують вагу минулих оцінок на теперішнє значення.

В контексті дисперсії гетероскедастичного процесу УАРУГ(p,q) має ковзне середнє та компоненти авторегресії. Високостаціонарний процес УАРУГ може на завжди бути слабостаціонарним, чи навпаки, у зв'язку з неоднозначною ідеєю стаціонарності по відношенню до узагальнених умовно гетероскедастичних процесів. Це робить більш складним виявлення та оцінку стаціонарності в таких типах систем [62, 63].

3.1.4 Методи оцінювання параметрів узагальнених лінійних моделей при неповних даних

Оцінювання параметрів УЛМ є серйозною задачею, що потребує ретельного огляду. Для порівняння методів, параметри були оцінені з використанням наступних алгоритмів: ітераційно-рекурсивно зважуваний метод найменших квадратів, алгоритм оптимізації Adam і метод Монте-Карло для ланцюгів Маркова [2].

Підхід ітераційно рекурентно зважуваних найменших квадратів або iteratively reweighted least squares (IRWLS) може бути застосований для задання гетероскедастичної моделі:

$$\tilde{\beta} = (X^T W X)^{-1} X^Y W y, \tag{3.40}$$

де y — цільова змінна;

$W = Diag[Var(y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2]^{-1}$ — діагональна матриця ваг;

X — матриця коваріат.

Використовується скоринг Фішера [64]:

$$\beta^{(t+1)} = (X^T W X)^{-1}(X^T W \beta^{(t)} + X^T A(y - \mu)) \tag{3.41}$$

де — $A = Diag[Var(y_i)(\frac{\partial \eta_i}{\partial \mu_i})]^{-1}$.

Між матрицями A та W існує наступний зв'язок:

$$A = W\left(\frac{\partial \eta}{\partial \mu}\right) = W\, Diag(\frac{\partial \eta_i}{\partial \mu_i}). \tag{3.42}$$

Звідси можна наступним чином записати вираз для оцінювання параметрів УЛМ:

$$\beta^{(t+1)} = (X^T W X)^{-1}(X^T W \beta^{(t)} + X^T W z), \tag{3.43}$$

де $z = \eta + (\frac{\partial \eta}{\partial \mu})(y - \mu)$.

На кожному етапі алгоритму:

1) нова функціональна змінна z та множина ваг W визначається за рахунок застосування поточної оцінки β;

2) для отримання нових значень, значення z оновлюється шляхом використання X та W.

Алгоритм адаптивної оцінки руху або Adam розширює алгоритм градієнтної оптимізації. Він був розроблений для пришвидшення оптимізаційної процедури з метою покращення ефективності алгоритму оптимізації. Це виконується за допомогою збільшення розміру кроку для кожного вхідного параметру під час оптимізації. Процедура пошуку, що ґрунтується на часткових похідних або градієнтах для кожної вхідної змінної, автоматично налаштовує кожний розмір кроку. Для вхідних величин це передбачає розрахунок першого та другого моменту градієнту як експоненційних спадних першого та другого моментів.

Для кожного параметру оптимізації, що представляється у вигляді m та v, у першу чергу модифікується вектор моментів та експоненціально зважена норма. Вони починаються з нульових значень.

Алгоритм складається з нижчеперелічених кроків.

Для поточного кроку розраховується градієнт цільової функції [65]:

$$g(k) = f(x(k-1)). \tag{3.44}$$

1. Потім використовується градієнт і гіперпараметр β_1 для оновлення першого моменту:

$$m(k) = \beta_1 * m(k-1) + (1-\beta_1) * g(k). \tag{3.45}$$

2. Квадрат градієнта і гіперпараметр β_2 застосовуються для оновлення другого моменту:

$$v(k) = \beta_2 * v(k-1) + (1-\beta_2) * g^2(k)\beta_2. \tag{3.46}$$

Так як їх початкові значення дорівнюють нулю, останні два значення зазвичай зміщені.

3. Наступний вираз використовується для оновлення першого та другого моменту:

$$\tilde{m}(k) = \frac{m(k)}{1-\beta_1^k}; \tag{3.47}$$

$$\tilde{v}(k) = \frac{v(k)}{1-\beta_2^k}. \tag{3.48}$$

4. Для визначення оптимальної точки використовується формула:

$$x(k) = x(k-1) - \alpha * \frac{\tilde{m}(k)}{\sqrt{\tilde{v}(k)}+\varepsilon}, \tag{3.49}$$

де α — гіперпараметр розміру кроку;

ε — мале значення, що гарантує неможливість виникнення помилки ділення на нуль.

Однією з переваг байєсівського переходу — його адаптивність або здатність повторно оцінювати параметри моделі, коли нові дані неперервно надходять.

Процедуру можна описати в цілому в наступному вигляді: параметри моделі можуть бути оцінені за допомогою даних на певному кроці i з використанням методу Монте-Карло для марківських ланцюгів (МКМЛ):

$$\tilde{\theta}_i = \tilde{\theta}_{i-1} + \alpha \Delta \tilde{\theta}(i), \tag{3.50}$$

де $\tilde{\theta}_i$ — оцінка деякого параметра θ моделі на кроці i;

α — ваговий фактор;

$\Delta \tilde{\theta}(i)$ — інкремент в значенні оцінки, що має бути обчислена.

Так як оцінка параметра на наступному кроці залежить лише від оцінки на попередньому кроці, критерій Маркова для застосування марківських ланцюгів з метою опису випадкових процесів є задовільним. Для обчислення зростання оцінки, дані з певного розподілу генеруються з використанням методу Монте-Карло. Звідси й загальний термін для методу Монте-Карло для марківських ланцюгів.

Як правило, байєсівський метод для побудови регресійної моделі задається наступним чином:

$$y_k \sim N(a_1 x_{k1} + \ldots + a_n x_{kn}, \sigma), \; k = 1, \ldots m, \tag{3.51}$$

де N — нормальний розподіл з відповідними параметрам.

Загалом, тип розподілу може бути обраний з врахуванням реального розподілу даних. Матрична форма може бути використана для іншого запису:

$$y \sim N(X\alpha, I),\ a \sim N(\alpha_0, K_{\alpha,0})\sigma_e, \tag{3.52}$$

де α_0 — вектор середніх значень на початковому кроці;

$K_{\alpha,0}$ — відповідна матриця коваріацій.

Початкова оцінка, відома як нульовий крок, буде перевизначена з урахуванням нової інформації та поступової адаптації.

Коли одночасно залежна змінна та компоненти, що оцінюються, є нормально розподіленими випадковими величинами, можливе застосування методу найменших квадратів (МНК). Але МКМЛ є зручним так як він дозволяє змінювати декілька параметрів розподілу без потреби повторного обчислення усіх матричних операцій. З сімейства нормальних, гамма та їх обернених розподілів, так звана властивість спряжених апріорних розподілів робить це можливим.

Регресійна модель записується наступним чином [2]:

$$y = a + bx + \epsilon. \tag{3.53}$$

Початкова оцінка вектору α_0 виражається у вигляді наступного виразу:

$$\alpha_0 = \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}, \tag{3.54}$$

а матриці коваріації:

$$K_{\alpha,0} = \begin{bmatrix} \sigma_{a,0}^2 & 0 \\ 0 & \sigma_{b,0}^2 \end{bmatrix}. \tag{3.55}$$

Формула для оновлених значень (t_1, s_1), (t_2, s_2) параметрів розподілу

може бути знайдено за допомогою застосування попередньо заданої властивості до кожної нової пари спостережень:

$$\alpha_1 = K_{\alpha,1}\left(\alpha_0 K_{\alpha,0}^{-1} + K_y^{-1}\mu_y\right); \tag{3.56}$$

$$K_{\alpha,1} = \left(K_{\alpha,0}^{-1} + K_y^{-1}\right)^{-1}; \tag{3.57}$$

$$\mu_y = \left[\frac{t_1 s_2 - t_2 s_1}{t_1 - t_2}, \frac{s_1 - s_2}{t_1 - t_2}\right]^T; \tag{3.58}$$

$$K_y = \begin{bmatrix} \left[\left(\frac{t_1}{t_1 - t_2}\right)^2 + \left(\frac{t_2}{t_1 - t_2}\right)^2\right]\sigma_e^2 & 0 \\ 0 & 2\left(\frac{1}{t_1 - t_2}\right)^2 \sigma_e^2 \end{bmatrix}. \tag{3.59}$$

Але у більшості випадків аналізу нестаціонарних процесів існує ситуація ймовірнісної невизначеності або невизначеності щодо розподілу, що означає, що неможливо гарантувати, що нормальний розподіл змінних та параметрів залишається незмінним, що робить неможливим узгоджене застосування властивості, згаданої раніше. Крім того, оскільки передумови для використання МНК були порушені, є технічно недоцільним продовжувати його використовувати. МКМЛ забезпечує засіб для того щоб апроксимувати невідомий розподіл за допомогою генерування даних для апостеріорного розподілу навіть за умови, коли тип шуканого розподілу невідомий.

3.2 Байєсівський підхід до оцінювання параметрів моделі

Маючи в наявності релевантні дані, байєсівський підхід може бути застосований для оцінювання параметрів моделі, θ, та пов’язаної з ними невизначеності. Це може бути взято до уваги під час розробки надійних моделей, які стійкі до невизначеності. Метод ґрунтується на теоремі Байєса, де дослідник розробляє апріорний розподіл ймовірностей для параметрів, $p(\theta)$, що оновлюється за ймовірністю спостереження даних x, $p(x|\theta)$, використовуючи

власне розуміння фізичної поведінки системи та математичних обмежень.

Це дає можливість оцінити $p(\theta|x)$ — апостеріорну ймовірність набору параметрів на основі спостережуваних даних [66]:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}. \tag{3.60}$$

3.2.1 Ітеративні методи Монте-Карло

Ефективним статистичним методом оцінювання апостеріорних розподілів є МКМЛ. Вибірки МКМЛ створюються за рахунок ланцюгів Маркова за допомогою стимуляції Монте-Карло. Відносна апостеріорна щільність поточного та запропонованого розташування визначає ймовірність переходу від поточної точки в ланцюзі до наступної запропонованої точки в просторі параметрів. Розподіл, що марківський ланцюг створює за допомогою вибірки в кінцевому рахунку стає стаціонарним та відображає справжній апостеріорний розподіл. Водночас, якщо апостеріорні оцінки МКМЛ застосовуються занадто рано до того, як алгоритм досягне свого апостеріорного розподілу, невизначеність може бути недостатньо якісно охарактеризована, що веде до систем, які переналаштовані або недостатньо налаштовані [67].

В результаті, створення діагностик для оцінення збіжностей пошукових алгоритмів МКМЛ став об’єктом великої кількості робіт. Ці діагностики зокрема зосереджені на знаходженні того, скільки початкових ітерацій має бути відкинуто як “burn-in”, так як вони не є репрезентативними для стаціонарного розподілу і далекі від апостеріорного режиму, як і те, скільки необхідно ітерацій для зупинки алгоритму, тому що ланцюг досяг свого стаціонарного розподілу.

Найпоширенішим алгоритмом МКМЛ є семплер Гіббса. Використовуючи поточні значення інших координат, семплер Гіббса створює послідовні величини з одновимірних повних умовних розподілів на основі довільної

початкової токи. Як результат, семплер Гіббса не створює незалежні та однаково розподілені послідовності довільних векторів. З більшою ймовірністю він створює ланцюг Маркова випадкових векторів, чий розподіл наближається до розподілу з щільністю f. Використовуючи вектори x_i, утворені марківськими ланцюгами, можна продемонструвати, що середнє $q(x_i)$ збігається до інтегралу вигляду $\int q(x)f(x)dx$ при звичайних критеріях регулярності. Зазвичай, вибір початкового значення грає доволі важливу роль. Це є гарною відправною точкою, якщо відоме приблизне розташування моди. У протилежному випадку, семплер працює до тих пір, поки статистичний тест не продемонструє, що марківський ланцюг досяг збіжності. Генерація вибірки згодом продовжується для створення векторного зразка та визначення середнього значення $q(x_i)$, відкидаючи перший сегмент ланцюга, часто відомий як “burn-in”. Обчислення довірчого інтервалу для результату інтеграції з використанням семплера Гіббса є можливим, але доволі складним процесом [68].

Одним з найбільш відомих методів МКМЛ, що дає можливість генерувати вибірку зі складних апостеріорних розподілів, є алгоритм Метрополіса-Гастінгса. Перед генерацією або пропозицією нових параметрів θ' із запропонованого розподілу, центрованого в поточному місці розташування, початково налаштований параметр θ_0 вибирається з попереднього розподілу, яким зазвичай виступає багатовимірний нормальний розподіл, для пропозиції нових налаштувань параметра. Даний процес описується, як гаусівська мутація.

Вираз (3.61) визначає ймовірність α що запропонований крок буде прийнятий [66]:

$$\alpha = min\left(1, \frac{p(\theta'|x)g(\theta_t|\theta')}{p(\theta_t|x)g(\theta'|\theta_t)}\right), \tag{3.61}$$

де $g(\theta'|\theta_t)$ — ймовірність того, що запропоновані параметри θ' на основі поточних параметрів дорівнює θ_t;

$g(\theta_t|\theta')$ — обернений вираз.

3.2.2 Неітеративні методи Монте-Карло

Для важливісного вибіркового моделювання пошук щільності вибірки за важливістю, також відомого як щільність пропозиції $g(x)$ є необхідним. У більшості застосувань, декілька характеристик для кількох маржинальних розподілів має бути знайдено. Тому ідеальну щільність значущості для одного $q(x)$ нереально знайти на практиці. У роботах стверджується, що в цьому випадку $g(x)$ має бути приблизно пропорційною до $f(x)$, але з більшими хвостами. Пріоритетний семплінг включає вибір векторів із запропонованої щільності $g(x)$ та оцінки інтегралу вигляду [66]:

$$\int q(x)w(x)g(x)dx, \tag{3.62}$$

де $w(x) = \frac{f(x)}{g(x)}$

В ролі функції ваги виступає $w(*)$, а вагою для x відповідно $w(x)$. Оскільки інтеграл для f невідомий, то має застосовуватися оцінка співвідношення $E_g(q(x)w(x)) \big/ E_g(g(x))$, як показано в [66] для її характеристик, замість традиційної оцінки вибіркової значущості $E_g(q(x)w(x))$. Очевидно, що ідеальні ваги мають бути приблизно постійними, або, інакше кажучи, мають певними чином змінюватися. Рекомендується використовувати так званий ефективний розмір вибірки в [66]:

$$n_e = \frac{\left(\sum w_i\right)^2}{\left(\sum w_i^2\right)} = \frac{n}{V(w_i)+1}. \tag{3.63}$$

Максимізація ефективного розміру вибірки рівносильно мінімізації дисперсії ваг, тобто:

$$n_e = \frac{n}{\int \frac{f^2(x)}{g(x)} dx}. \tag{3.64}$$

Метод повторної вибірки, що має назву методу bootstrap був розроблений для оцінки точності статистичного оцінювання та побудови висновків щодо невідомих характеристик генеральної сукупності, як середнє, медіана, дисперсія і довірчий інтервал. За рахунок його легкості у використанні та ефективності у наданні точних оцінок, він широко застосовується в дослідженнях. Оскільки вибірковий розподіл не завжди легко отримати в практичних застосуваннях, bootstrap розподіл може точно відтворити вибірковий розподіл для будь-якої статистики, що цікавить.

Метод bootstrap реалізовується за допомогою bootstrap-вибірок з використання параметричних і непараметричних моделей, а потім розрахунок цільової статистики на основі кожної bootstrap-вибірки. Можливо сформулювати висновки щодо релевантної статистики за допомогою застосування емпіричного розподілу результатів як розподіл, що замінює вибірковий розподіл.

Для початку потрібно ввести декілька позначень для застосування методу Bootstrap з використанням інтеграції Монте-Карло. Припускається, що незалежні та однаково розподілені випадкові величини $X_1;X_2;...X_n$ мають носіїв на $[a,b]$ і відповідають розподілу ймовірностей f [69]. Також припускається, що спостереження, які корелюють з цими випадковими величинами, є $x_1;x_2;...x_n$. Крім того, нехай $\theta = \int_a^b g(x)dx$ являє собою цільовий інтеграл. Такі етапи, як представлено нижче, спрощують демонстрацію методології методу bootstrap з використання методу Монте-Карло.

1. Генеруються n спостережень з ймовірнісного розподілу f.

2. Знаходяться $g(x_i)$ для всіх $i = 1,2,...,n$.

3. Обчислюється

$$\hat{\theta}^* = \frac{b-a}{n}\sum_{i=1}^{n} g(x_i). \tag{3.65}$$

4. Виконуються кроки 1,2 та 3 B разів; це веде до $\hat{\theta}^{*1}, \hat{\theta}^{*2},...,\hat{\theta}^{*B}$.

Для кращої роботи алгоритму, в якості значення для B рекомендується використовувати велике число, наприклад $B = 1000$. Для виведення bootstrap оцінки θ , обчислюється середнє $\hat{\theta}^{*1}, \hat{\theta}^{*2},...,\hat{\theta}^{*B}$ і з метою надання оцінки стандартної помилки

$$\overline{\hat{\theta}^*_{boot}} = \frac{1}{B}\sum_{j=1}^{B} \hat{\theta}^{*j}, \tag{3.66}$$

розраховується стандартне відхилення $\hat{\theta}^{*1}, \hat{\theta}^{*2},...,\hat{\theta}^{*B}$:

$$SE\left(\overline{\hat{\theta}^*_{boot}}\right) = \sqrt{\frac{\frac{1}{B}\sum_{j=1}^{B}\left(\hat{\theta}^{*j}\right)^2 - \frac{\left(\sum_{j=1}^{B}\hat{\theta}^{*j}\right)^2}{B}}{B-1}}. \tag{3.67}$$

Обирається $\left(\frac{\alpha}{2}\right)$ та $\left(1-\frac{\alpha}{2}\right)$ впорядковані значення, де перше виконує роль нижньої межі, а останнє — верхньої межі після сортування значень для довірчого інтервалу $(1-\alpha)\%$ процентиля від найменшого до найбільшого. Це можна записати наступним чином:

$$\theta \in \left(\hat{\theta}^*_{\left(\frac{\alpha}{2}\right)}, \hat{\theta}^*_{\left(1-\frac{\alpha}{2}\right)}\right). \tag{3.68}$$

Наступний вираз використовується після обчислення $\overline{\hat{\theta}^*_{boot}}$ та $SE\left(\overline{\hat{\theta}^*_{boot}}\right)$ для $(1-\alpha)\,\%$ нормального довірчого інтервалу:

$$\theta \in \left(\overline{\hat{\theta}^*_{boot}} - Z_{(1-\alpha)} SE\left(\overline{\hat{\theta}^*_{boot}}\right), \overline{\hat{\theta}^*_{boot}} + Z_{(1-\alpha)} SE\left(\overline{\hat{\theta}^*_{boot}}\right)\right), \tag{3.69}$$

де $Z_{(1-\alpha)}$ — $(1-\alpha)$ процентиль стандартного Гаусівського розподілу.

3.3 Порівняння точності та швидкодії методів і приклади їх застосування

Попередня обробка даних, що включає в собі заповнення пропусків, нормалізація до визначеного діапазону, перетворення категоріальних даних і фільтрація, що дозволяє враховувати як викиди так і екстремальні значення, є невід’ємною складовою в будь-якій задачі інтелектуального аналізу даних. Водночас є недоречним робити остаточні судження щодо результативності моделей апріорі на етапах фільтрації та заповнення пропусків даних. Крім того є можливість дати відповідь на питання щодо коректності застосування нормалізації та денормалізації та на питання щодо того, чи використовуються коректні атрибути у подальших кроках тесту Хі-квадрат та методу RFE.

3.3.1 Оцінювання якості моделей

Дослідження ряду залишків або похибок моделі може застосовуватися для визначення того, чи є модель регресії адекватною. У цьому випадку значення, що оцінюються, знаходяться за допомогою заміни дійсних значень множини елементів, що включені в модель. Ряд залишків досліджується на наявність характеристик стохастичної компоненти часового ряду, що зазвичай має на увазі гаусовість закону розподілу, близькість математичного сподівання до нульового значення, стохастичність похибок та відсутність автокореляції.

Використовують певні критерії адекватності, що застосовуються для оцінки адекватності моделі.

Сума квадратів залишків (RSS). Представляє собою вираз у вигляді суми квадратів залишків, що є сумою квадратів різниць між змодельованими та дійсними значеннями пояснювальної змінної як для часу так й ідентифікації:

$$RSS = \sum_{k=1}^{n} \left(y(k) - \tilde{y}(k)\right)^2. \quad (3.70)$$

Чим менше значення, тим більш адекватна модель.

Коефіцієнт детермінації. Являє собою статистичну міру інформативності моделі по відношенню до даних. Даний показник відображає ступінь до якої модель та присутні спостереження збігаються:

$$R^2 = 1 - \frac{var(\tilde{y})}{var(y)}. \quad (3.71)$$

Модель є адекватною, якщо значення даного критерію близьке до одиниці.

Критерій Дарбіна-Вотсона (DW). Застосовується у вигляді виразу для перевірки автокореляції першого порядку ряду, що досліджується. Зазвичай він використовується для залишків моделі регресії та часових рядів:

$$DW = \frac{\sum_{k=2}^{K} (e_k - e_{k-1})^2}{\sum_{k=1}^{K} e_k^2} \approx 2(1 - \rho_1), \tag{3.72}$$

де ρ_1 — коефіцієнт автокореляції першого порядку.

За умови відсутності автокореляції DW дорівнює 2, у протилежному випадку DW набуває нульового значення. Якщо автокореляція менше нуля, то DW дорівнює 4.

$$\begin{cases} DW = 2, if \ \ \rho_1 = 0 \\ DW = 0, if \ \ \rho_1 = 1 \\ DW = 4, if \ \ \rho_1 = -1 \end{cases}. \tag{3.73}$$

Інформаційний критерій Акайке (AIC). Цей показник вимірює відносну якість статистичних моделей для певного набору даних. *AIC* оцінює якість кожної моделі по відношення до інших моделей за наявності множини моделей для даних. Базуючись на теорії інформації, *AIC* описує відносні оцінки втраченої інформації, коли певна модель застосовується для відображення процесу генерації даних. В кінцевому результаті, це демонструє компромісний варіант між складністю моделі та ступенем узгодженості. Висунемо припущення, що є статистична модель деяких даних. Припустимо, що k виконує роль числа параметрів моделі, що були оцінені, а L — максимум функції правдоподібності. Звідси *AIC* знаходиться за допомогою наступної формули:

$$AIC = 2k - 2ln(L). \tag{3.74}$$

Менше значення вказує на те, що дана модель є більш адекватною.

Байєсівський інформаційний критерій (BIC) або інформаційний критерій Шварца. Представляє собою статистичний критерій для вибору моделі зі скінченної множини моделей. Він нерозривно пов'язаний за інформаційним критерієм Акайке і частково ґрунтується на функції правдоподібності:

$$BIC = k\,ln\,(n) - 2\,ln(\tilde{L}), \tag{3.75}$$

де $\tilde{L}$ — максимум функції правдоподібності моделі;

n — розмірність датасету, тобто кількість спостережень;

k — число оцінених параметрів моделі.

3.3.2 Оцінювання якості прогнозів

Об'єктивне оцінювання якості прогнозу є найбільш пріоритетним етапом в процесі прогнозування. Якість прогнозованих значень має бути оцінено з застосуванням числа статистичних критеріїв, так як вони представляють собою випадкові змінні. Структуру поділу даних, що застосовується для оцінки моделі та аналізу якості прогнозу описано на рисунку 3.2.

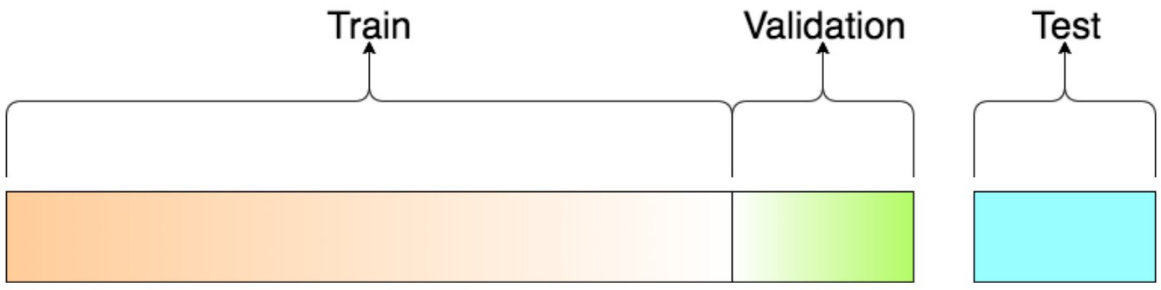

Рисунок 3.2 — Ділення набору даних перед навчанням і прогнозом моделі[3]

[3] Джерело зображення: https://i.sstatic.net/3tHZF.png.

Наявну вибірку даних коректно ділити на навчальну, валідаційну і тестову. Дійсність однокрокового прогнозу на даному сегменті ряду може бути визначена з метою оцінки параметрів моделі процесу і застосування історичного прогнозу.

В науковій літературі, прогноз на валідаційній вибірці даних інколи ще називають прогнозом ex post. Значно більший набір ряду має бути використаний для оцінки параметрів моделі при дослідженні рядів малої розмірності. Прогнозом ex ante називають прогнозні значення на тестовій вибірці.

Адекватність прогнозів як правило аналізується за рахунок застосування допоміжних метрик прогнозу. Так, показник MSE нерозривно пов'язаний з розмірністю даних, тому його поодиноке використання для надання оцінки якості прогнозу є недостатнім.

Дійсність та точність є необхідними складовими у будь-якому прогнозі. Тут застосовується множина стандартів, методів та підходів для аналізу якості прогнозу.

На практиці зазвичай використовують метрики якості прогнозу.

Середньоквадратична похибка (MSE). Кожен залишок регресії підносить до квадрату, додається, і в кінці результат ділить на загальне число відхилень для визначення даного показника:

$$MSE = \frac{\sum_{k=1}^{m} (y(k) - \tilde{y}(k))^2}{m}. \quad (3.76)$$

Корінь середньоквадратичної похибки (RMSE). Розраховується шляхом взяття квадратного кореня від показника MSE:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{m} (y(k) - \tilde{y}(k))^2}{m}}. \quad (3.77)$$

Коефіцієнт нерівності Тейла. За попереднім заданим описом, $U \in [0,1]$,коефіцієнт нерівності Тейла є одним із найважливіших показників якості як прогнозу, так і моделі. Вираз для розрахунку U свідчить, що якщо $U = 1$, то модель має непридатні характеристики прогнозу:

$$U = \frac{\sqrt{\frac{\sum_{k=1}^{M} (y(k) - \tilde{y}(k))^2}{M}}}{\sqrt{\frac{\sum_{k=1}^{M} y^2(k)}{M}} + \sqrt{\frac{\sum_{k=1}^{M} \tilde{y}^2(k)}{M}}}. \quad (3.78)$$

Модель оцінюється як ідеальна, коли $U = 0$, маючи на увазі, що дійсні та спрогнозовані дані ряду подібні. Інакше кажучи, це дає можливість з'ясувати чи модель або метод є доречним в теоретичному сенсі для оцінки прогнозу.

Середня абсолютна похибка у відсотках (MAPE). Представляє собою середнє значення в абсолютному контексті у відсотках по відношенню до дійсного значення показника:

$$MAPE = \frac{1}{M} \sum_{k=1}^{M} \frac{|y(k) - \tilde{y}(k)|}{|y(k)|} * 100\% = \frac{1}{M} \sum_{k=1}^{M} \frac{|\varepsilon(k)|}{|y(k)|} * 100\%, \quad (3.79)$$

У випадку задачі прогнозування на певну кількість кроків вперед, наприклад t, вираз набуває наступного вигляду:

$$MAPE = \frac{1}{t} \sum_{j=1}^{t} \frac{|y(k+j) - \tilde{y}(k+j,k)|}{|y(k+j)|} * 100\% = \frac{1}{t} \sum_{k=1}^{t} \frac{|\varepsilon(k+j)|}{|y(k+j)|} * 100\%. \quad (3.80)$$

Ця метрика в загалом застосовується для аналізу точності пронозу гетерогенних об'єктів, тобто прогнозних процесів, так як вона описує характеристики відносної якості пронозу. З іншого боку, у зв'язку з тим, що відносна міра легше сприймається як і дослідниками так і звичайними

користувачами, як правило це є доцільним при проведенні порівняльного аналізу якості прогнозу одного процесу різноманітними підходами.

Значення $MAPE$ збігатиметься до нескінченності якщо $y(k)$ та $y(k+j)$ рухаються до нульового значення в (3.79) та (3.80). При застосування даного критерію якості прогнозу це потрібно враховувати. Нульові значення $y(k)$ та $y(k+j)$ повинні бути відкинуті за допомогою належної модифікації значення M або t з метою розрахунку даного показника при певних умовах. Хоча цей метод може і не задовольняти усі критерії статистичного аналізу даних, він дійсно надає більш детальну та апроксимовану оцінку якості прогнозу.

3.3.3 Прийняття рішень на основі отриманих результатів

У цьому досліді прийнято рішення використовувати імітаційне моделювання для випадкового генерування цільової змінної та страхових індикаторів, так як страхові дані не завжди доступні всьому суспільству. Такі характеристики задають структуру даних [3]:

— **вік:** числова змінна, яка генерувалася в діапазоні між 18 та 64;

— **стать:** категорійна строкова змінна;

— **індекс маси тіла:** числова змінна, яка генерувалася за допомогою нормального розподілу;

— **кількість дітей:** числова змінна, яка генерувалася в діапазоні від 0 до 5;

— **статус курця:** категорійна строкова змінна;

— **регіон:** категорійна строкова змінна, що генерувалася з вибірки: “east”, “south”, “west”, “center”, “north”;

— **виплати:** числова змінна.

До останньої величини, яка є цільовою, застосовувалися наступні закони розподілу з відповідними функціями зв’язку:

— нормальний розподіл з відомою дисперсією та логарифмічною функцією зв’язку;

— експоненційний розподіл з тотожною функцією зв’язку;

— розподіл Парето з відомим параметром масштабування x_m та функцією зв’язку виду $f(x) = -1 - x$.

Гаусівський шум зі змінною дисперсією, що мала вид лінійної функції, додавався до цільової змінної.

Приклад згенерованих даних наведено на рисунку 3.3.

	age	sex	smoker	region	bmi	children	charges
0	22	female	no	west	33.136	2	5048.41
1	43	female	yes	center	21.386	1	5021.11
2	21	male	no	center	22.646	2	5041.95
3	37	female	yes	center	30.546	2	4858.78
4	41	male	no	center	31.623	2	4954.93

Рисунок 3.3 — Приклад вибірки згенерованих даних, що використовувався при дослідженні

Використовується адаптація зі сторони даних і зі сторони моделей. У першому випадку розширюємо набір даних і працюємо з ними. Якщо нові дані надходять в СППР, то набір даних розширюється і нові дані включаються в аналіз. У другому випадку використовуємо декілька моделей паралельно і на виході обираємо модельна за якою отримали більш точний результат.

Окрім імітаційних даних ці експерименти були зроблені і на реальних даних, і ми їх порівнювали, щоб перевірити, що і на імітаційних, і на реальних даних маємо однакові результати.

Перед оцінюванням параметрів виконувалася попередня обробка даних. Для всіх числових змінних застосовувалася стандартизація, тобто від кожної змінної віднімалося вибіркове середнє і ділилося на стандартне відхилення. Для категорійних змінних застосовувався підхід Label Encoder.

Після застосування фільтру Калмана дійсні страхові значення були порівняні з відфільтрованими значеннями.

Графіки результатів застосування фільтру Калмана до цільових змінних з різними розподілами представлено на рисунках 3.4-3.6.

Наступні критерії оцінки прогнозування використовувалися для аналізу якості моделей: MSE, RMSE та MAE.

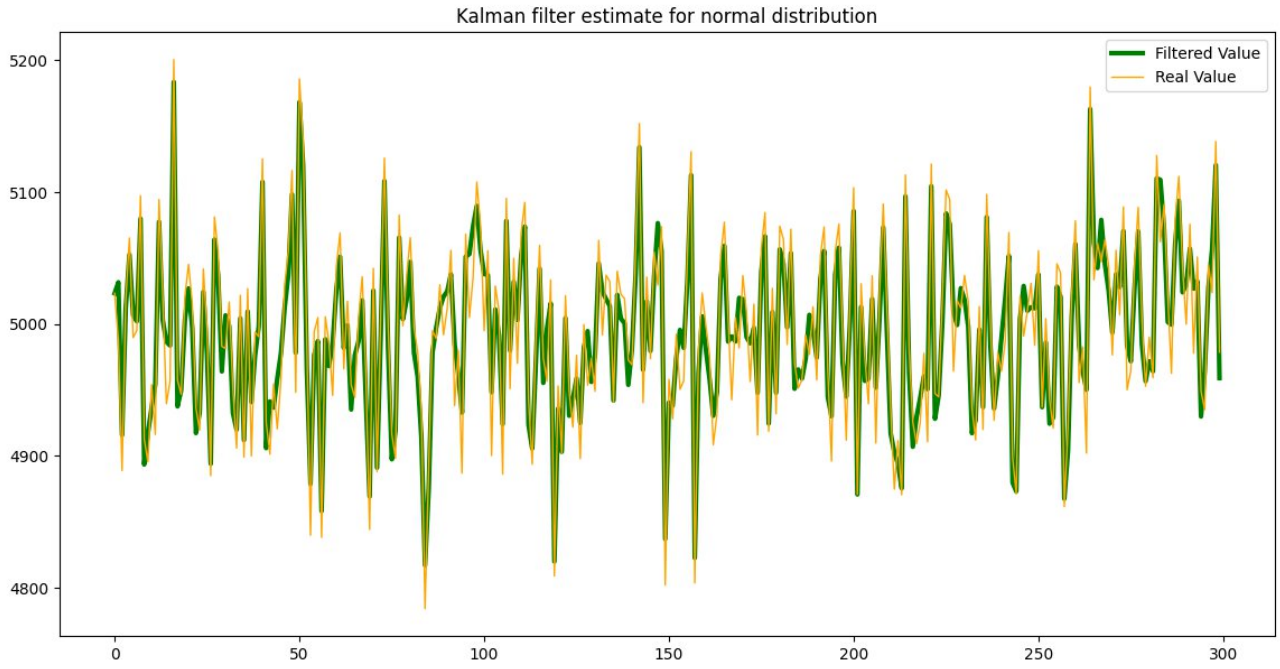

Рисунок 3.4 — Результат застосування фільтру Калмана до виплат, що мають нормальний розподіл

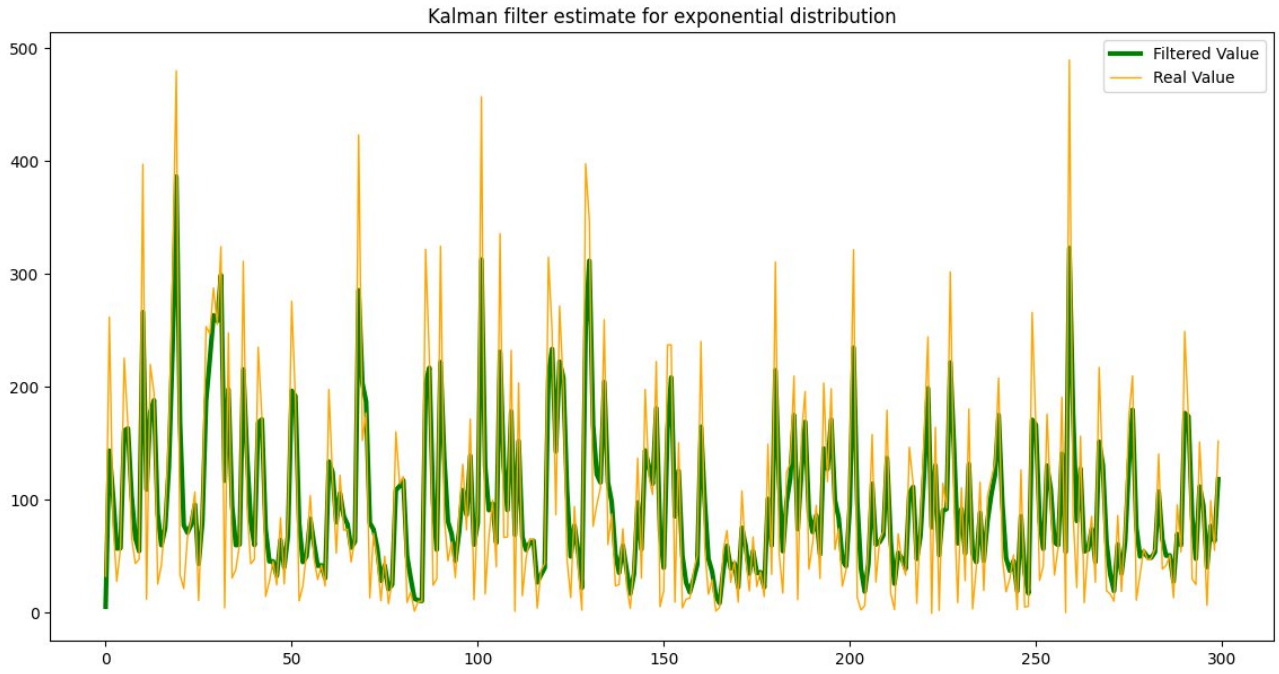

Рисунок 3.5 — Результат застосування фільтру Калмана до виплат, що мають експоненційний розподіл

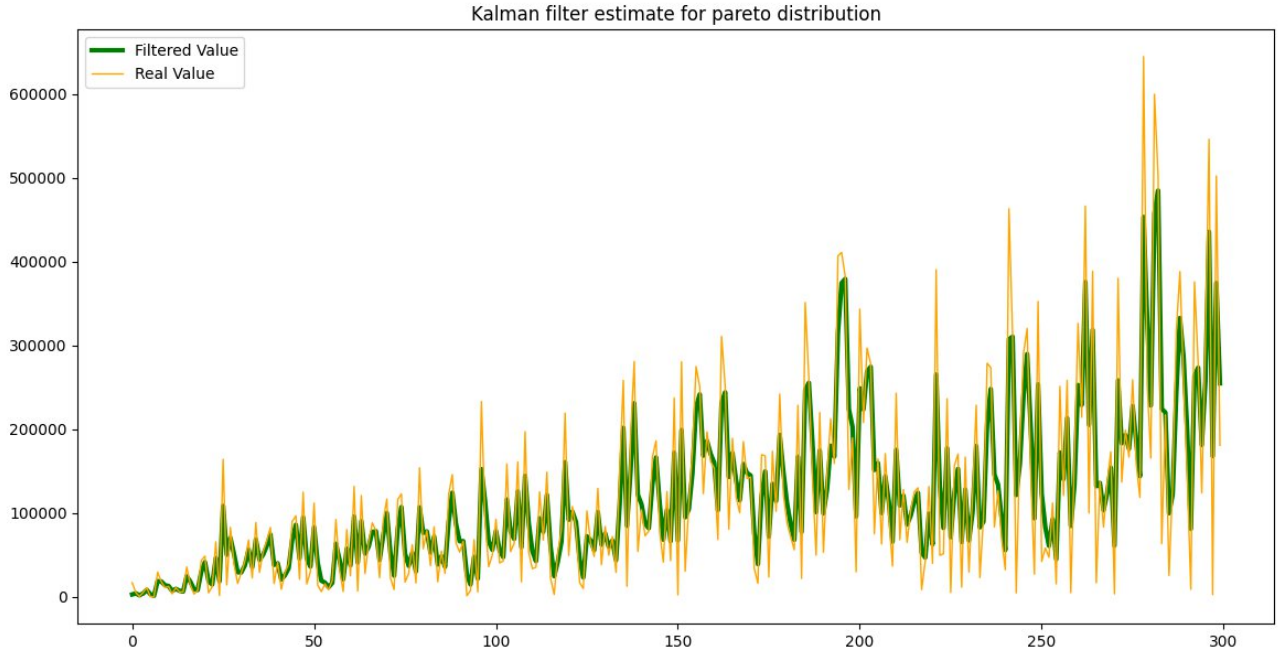

Рисунок 3.6 — Результат застосування фільтру Калмана до виплат, що мають розподіл Парето

Результати оцінювання параметрів УЛМ з використанням трьох підходів (IRWLS, Adam та МКМЛ) з та без використання фільтра Калмана для трьох запропонованих законів розподілу та заданих функцій зв'язку наведено в таблицях 3.1-3.6.

Таблиця 3.1 — Результати побудови УЛМ для виплат, що мають Гаусівський розподіл з логарифмічною функцією зв'язку без використання фільтру Калмана

Метрика	*МКМЛ*	*ADAM*	*IRWLS*
MSE	4408.68	684.63	2457.11
RMSE	64.35	25.1	48.4
MAE	48.76	23.76	48.1

Таблиця 3.2 — Результати побудови УЛМ для виплат, що мають Гаусівський розподіл з логарифмічною функцією зв'язку з використанням фільтру Калмана

Метрика	*МКМЛ*	*ADAM*	*IRWLS*
MSE	165.33	74.5	218.68
RMSE	12.85	8.63	14.79
MAE	3.957	5.65	11.77

Таблиця 3.3 — Результати побудови УЛМ для виплат, що мають розподіл Парето з від'ємною лінійною функцією зв'язку без фільтру Калмана

Метрика	*МКМЛ*	*ADAM*	*IRWLS*
MSE	52724.54	88185.03	638115.4
RMSE	228.51	286.65	797.61
MAE	153.03	205.31	771.25

Таблиця 3.4 — Результати побудови УЛМ для виплат, що мають розподіл Парето з від'ємно лінійною функцією зв'язку з використанням фільтру Калмана

Метрика	*МКМЛ*	*ADAM*	*IRWLS*
MSE	7858.747	16808.82	22900.647
RMSE	88.65	129.65	151.33
MAE	63.513	112.12	128.532

Таблиця 3.5 — Результати побудови УЛМ для виплат, що мають експоненціальний розподіл з тотожною функцією зв'язку без використання фільтру Калмана

Метрика	*МКМЛ*	*ADAM*	*IRWLS*
MSE	211.38	147.9	288.51
RMSE	14.47	12.18	16.7
MAE	11.05	3.63	13.7

Таблиця 3.6 — Результати побудови УЛМ для виплат, що мають експоненціальний розподіл з тотожною функцією зв'язку з використанням фільтру Калмана

Метрика	*МКМЛ*	*ADAM*	*IRWLS*
MSE	78.21	54.723	107.478
RMSE	8.84	7.397	10.367
MAE	4.1	2.34	7.147

Результати побудови УЛМ для трьох вищезгаданих випадків наочно демонструють, що підхід Adam в цілому показує прийнятні результати. Метод МКМЛ також показує задовільні результати в розподілі Парето. З результату стає зрозумілим, що застосування фільтру Калмана для початкової обробки даних та налаштування моделі покращує результати прогнозних метрик.

Висновки до розділу 3

Невизначеність є невід'ємною частиною при дослідженні будь-яких процесів. Її неможливо усунути повністю, але завжди можна мінімізувати. До них зазвичай відносять методи попередньої обробки числових даних (стандартизація і нормалізація), категорійних даних (Label encoding, One-hot encoding, Dummy encoding) та вибір найважливіших ознак (RFE, тест Хі-квадрат). Також при дослідженні потрібно враховувати ситуації відсутності даних та їх зашумлення. Окремим випадком невизначеності є статистична, ефективним засобом проти якої є методи фільтрації, як оптимальний фільтр Калмана.

Під час дослідження для оцінювання параметрів УЛМ застосовувався рекурентний ітераційно-зважувальний метод найменших квадратів, метод оптимізації Adam і метод Монте-Карло для ланцюгів Маркова.

У зв'язку з відсутністю в публічному доступі страхових даних, було прийнято рішення застосувати імітаційне моделювання. Виплати, що виконували роль цільової змінної, мали нормальний розподіл, експоненційний і розподіл Парето. В якості УЛМ використовувався нормальний розподіл з відомим стандартним відхиленням з логарифмічною функцією зв'язку, експоненційний розподіл з тотожною функцією зв'язку і розподіл Парето з відомим параметром масштабування і від'ємною лінійною функцією зв'язку. Розглядалися обидва із застосуванням та без фільтра Калмана. Результати значень метрик прогнозу вказують, що метод Adam показує в цілому непогані результати. Крім того метод Монте-Карло продемонстрував задовільні результати і для випадку розподілу Парето.

При розробці СППР було прийнято рішення у подальшому використовувати метод Монте-Карло для ланцюгів Маркова для оцінювання параметрів УЛМ. Це обґрунтовується тим, що він передусім дає можливість оцінювати параметри для різних типів розподілів. Алгоритм разом з цим дає можливість генерувати дані з апріорі заданим розподілом. Крім того даний

підхід можна застосовувати до оцінок параметрів методів Байєсівської фільтрації, як обробка даних і параметрів ймовірнісними алгоритмами фільтрації. Насамкінець, метод Монте-Карло для марківських ланцюгів є таким, що легко піддається адаптації до абсолютно нових даних.

РОЗДІЛ 4 СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

У сучасних умовах стрімкого зростання обсягів даних та ускладнення процесів управління в економічних, фінансових і технічних системах, ключову роль відіграють системи підтримки прийняття рішень (СППР). Вони забезпечують можливість аналітичної обробки інформації, формування моделей, оцінювання ризиків та вибору оптимальних альтернатив у ситуаціях невизначеності. Особливо актуальним це є для сфер, де важливо комбінувати як кількісні дані, так і знання експертів, — зокрема, в актуарній практиці та страхуванні.

СППР являють собою комплексні програмні та інформаційні системи, що поєднують математичні моделі, необхідні обчислювальні процедури, бази даних, засоби візуалізації та інструменти взаємодії з користувачем. Залежно від домінуючої концепції побудови, вони можуть ґрунтуватися на інформаційних підходах, технологіях штучного інтелекту або інструментальних програмних засобах. Їх архітектура зазвичай включає підсистеми для керування даними, моделювання, генерування результатів та представлення аналітичної інформації, що дозволяє забезпечити гнучкість і адаптивність до потреб користувача.

У межах цього розділу розглянуто побудову СППР для аналізу актуарних ризиків на основі фактичних страхових даних. Особливу увагу приділено інтеграції сучасних методів математичного моделювання — узагальнених лінійних моделей та мереж Байєса, які дозволяють комплексно оцінювати втрати та їх імовірності. Це забезпечує підвищення точності прогнозування страхових виплат і дає можливість отримувати інтерпретовані результати для підтримки управлінських рішень. Також у розділі представлено архітектуру розробленої СППР, технологічні інструменти, використані для її реалізації, а також приклади функціонування інтерфейсу користувача та візуалізації даних. Проведено порівняльний аналіз різних моделей, продемонстровано приклади обчислення ризику для страхових позовів та зроблено висновки щодо

ефективності запропонованого підходу. Окремо окреслено перспективи розвитку СППР у страхуванні, включно з переходом до хмарних архітектур, використанням REST API та інтеграцією сучасних платформ AWS, GCP та Azure.

4.1 Архітектура системи підтримки прийняття рішень

Архітектура розробленої системи підтримки прийняття рішень (СППР) побудована таким чином, щоб забезпечити повний цикл обробки страхових даних — від завантаження та попередньої обробки до побудови моделей і генерації підсумкових результатів для користувача. На рисунку 4.1 представлено структурну схему СППР для аналізу актуарних ризиків, що відображає взаємодію основних її компонентів.

Складники розробленого програмного забезпечення виконують такі функції:

— мовна система (LS або Language System): створення запитів для ретроспективного аналізу, представлення результатів та вибір технік обробки даних;

— система обробки даних і генерування результатів (DPSRG або Data Processing System and Result Generation): вибір та виконання алгоритму, вибір та розрахунок критеріїв, діагностичні сповіщення, і вибір та створення презентації результатів;

— система подання результатів (RPS або Result Providing System): графічний та текстовий формат;

— база даних (DB або Database): включає в себе дані, моделі, параметри моделей, алгоритм оцінювання параметрів моделей та прогнозу і метрики якості прогнозу.

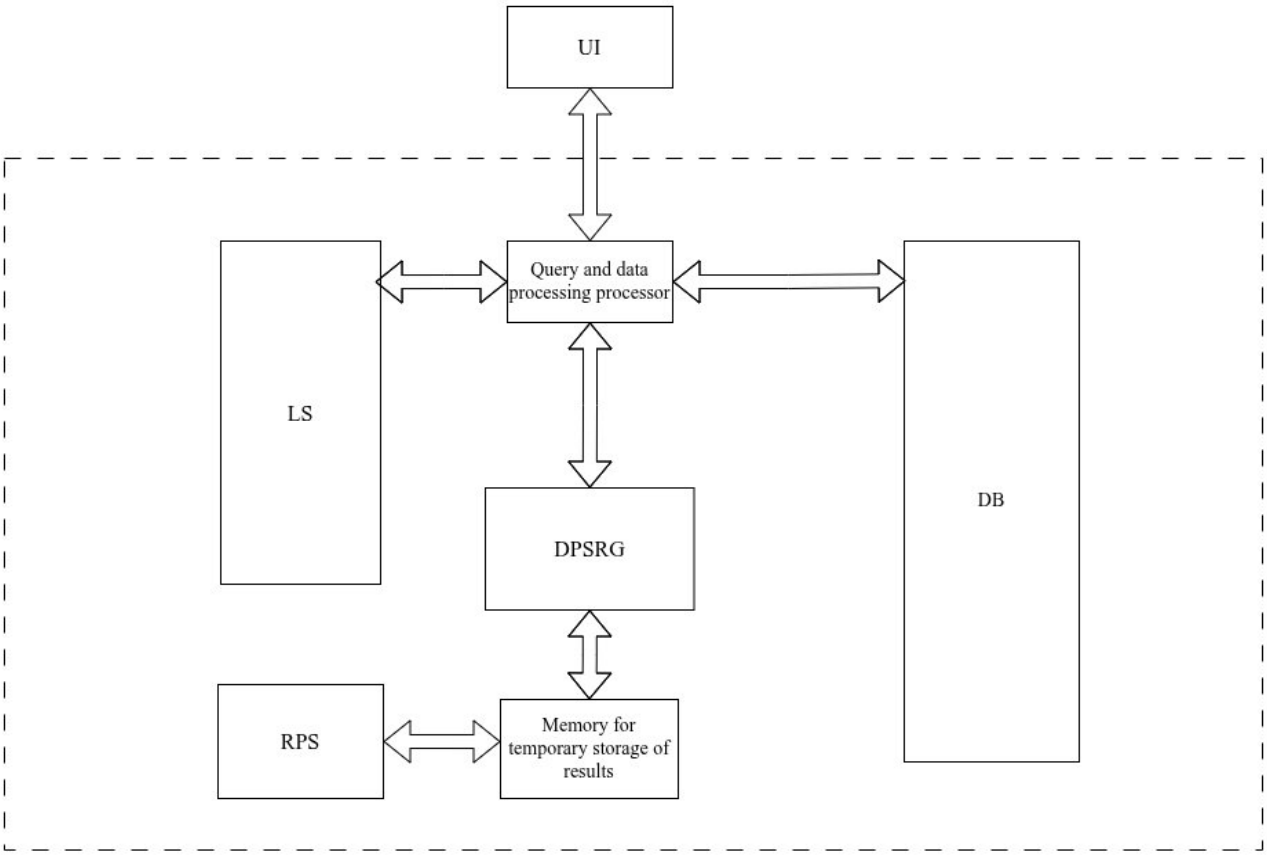

Рисунок 4.1 — Архітектура СППР

4.1.1 Класичні підходи до побудови системи підтримки прийняття рішень

На практиці при побудові СППР виділяють такі класичні підходи до розробки [34]:

— інформаційний підхід;

— підхід, що ґрунтується на знаннях;

— інструментальний підхід.

Інформаційний підхід: згідно цього підходу, СППР класифікують як інформаційну систему, головна задача якої полягає у застосуванні інструментів інформаційної технології з метою вдосконалення середовища керування діяльності персоналу, тобто покращення саме середовища, а не забезпечення

необхідної інформації у заданий час. Дві моделі СППР, “Спрага” та еволюціонуюча модель були запропоновані в якості параметрів даного підходу.

Інтерфейс “користувач-система”, база даних і база моделей являють собою три першочергові складові моделі СППР “Спрага”. Спілкування з кожною базою даних може бути реалізовано шляхом застосування інтерфейсу “користувач-система”. Система управління базою даних або СУБД, система управління базою моделей або СУБМ і створення та керування діалогами є прикладами інструментів програмного забезпечення. Інтерфейс має бути здатний до управління декількома стилями діалогів, змінювати діалоговий стиль за запитом користувача, забезпечувати дані в різних форматах та типах і пропонувати гнучку допомогу користувачам [34].

Схема структури СППР “Спрага” зображена на рисунку 4.2.

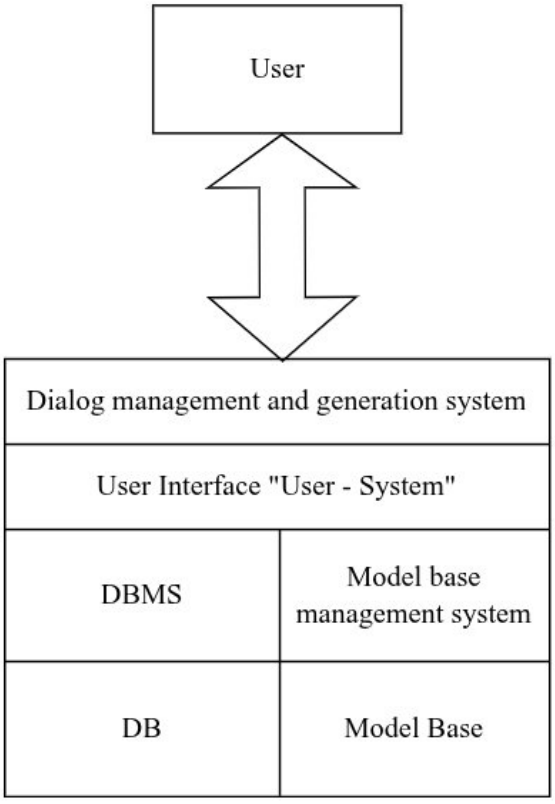

Рисунок 4.2 — Схема побудови СППР “Спрага”

І кількісні, і якісні дані з багатьох різних джерел включаються в базу даних СППР.

Створення і зберігання бази даних повинно забезпечувати такі функції:

1) здатність інтегрувати різні джерела інформації через їх процес вилучення даних;

2) представлення логічної структури в термінах, зрозумілих користувачу;

3) забезпечення всеохоплювального набору послуг щодо керування даних.

Зокрема, основа моделі має пропонувати гнучкість моделювання за рахунок використання апріорі створених модельних блоків та підпрограми моделі. Керування моделями пропонує такі функції:

1) здатність будувати нові моделі легко і зрозуміло;

2) здатність каталогізувати та обслуговувати різноманітні моделі, що підтримують усі рівні управління;

3) здатність поєднувати моделі з певними базами даних.

Еволюціонуюча СППР представляє собою розширену версію СППР “Спрага”. Ця система включає в собі текстову базу даних, базу даних правил, звичайну базу даних, базу даних моделей та інтерфейс. Внаслідок цього вони стають більш функціональними. Можна використовувати як структуровану інформацію, як правила відображення знань та евристичні методи, і менш структуровану, як тексти простою мовою, в інформаційній базі даних СППР [34].

Схема структури еволюціонуючої СППР наведена на рисунку 4.3.

Підхід, що ґрунтується на знаннях: одним з потенційних напрямків розробки СППР є комбінація технологій підтримки рішення та технології штучного інтелекту. Водночас в контексті класифікації СППР, є доречним враховувати модель СППР, що ґрунтується на знаннях.

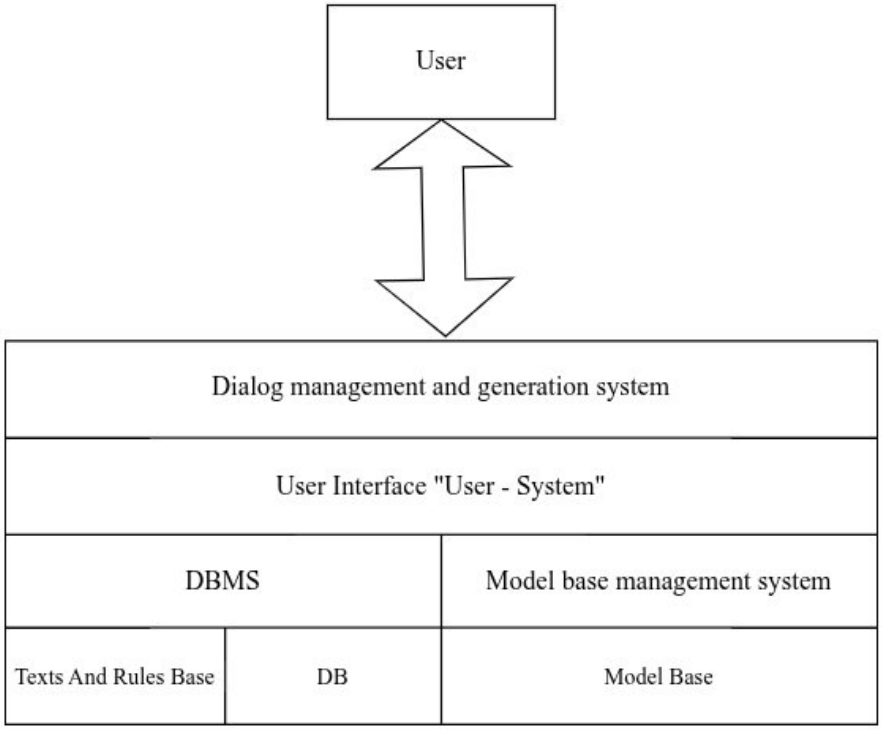

Рисунок 4.3 — Схема побудови еволюціонуючої СППР

Елементи штучного інтелекту, а саме застосування звичайної мови для спілкування з системою, методологія експертних систем, інженерія знань та комп'ютерні мови штучного інтелекту проявили себе у трьох базових складових СППР: база даних і СУБД, база моделей і система управління базою моделей, користувацький інтерфейс. Але існують концепції створення СППР, в яких база знань СППР виконує роль одного з впливових чинників. Відмінною ознакою СППР, що ґрунтується на знаннях є явне виділення нового аспекту підтримки рішень — здібність розуміти задачу, тобто вміння приймати користувацький запит, збирати релевантні дані та генерувати звіт [34].

Схема структури СППР, що ґрунтується на знаннях, наведено на рисунку 4.4.

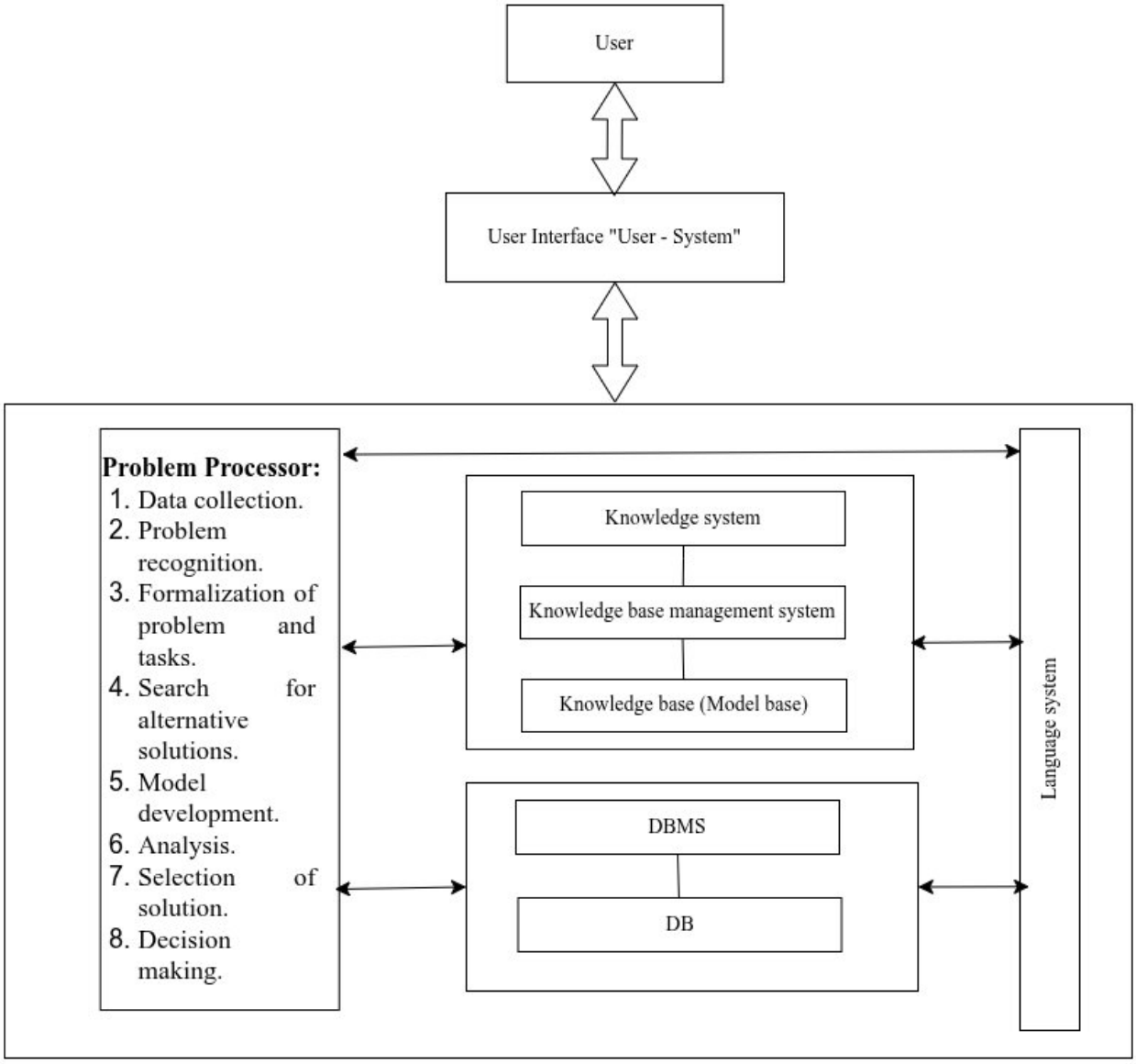

Рисунок 4.4 — Схема побудови СППР, що ґрунтується на знаннях

Мовна система (Language system), система знань (база даних, система управління базою даних, база знань (Knowledge base) і система управління базою знань (Knowledge base management system) і система обробки або рішення проблеми (Problem processor) являють собою три взаємопов’язані складові системи.

Користувач і кожна частина комп’ютерної системи поєднані за допомогою мовної системи. Використовуючи синтаксичні та семантичні засоби, надані мовною системою, вона допомагає користувачу у формулюванні

проблеми та керуванням процесу розв’язання. Інформації щодо предмету зберігається в системі знань.

Дані системи розрізняються в залежності від того, як представляються дані та моделі формалізації знань, що використовуються, що включає в себе, фрейми, ієрархічні структури, графи, семантичні мережі, предикатна арифметика, тощо.

Мовна система та система знань пов’язані за допомогою системи вирішення проблеми. Збір інформації, розробка моделі, аналіз та інші функції гарантуються за допомогою проблемного процесору. Для генерування потрібних даних для підтримки альтернатив вибору, він приймає на вхід опис проблеми, виконує її згідно синтаксису мовної системи і використовує знання у відповідності до правил, що встановлені в системі знань.

Виконуючи функцію динамічної частини СППР, проблемний процесор імітує або демонструє, як будь-яка особа підійшла б до певної проблеми. В кінцевому результаті, він має бути здатний включати інформацію для користувача через мовну систему та систему знань.

Проблемний процесор, що застосовує математичні моделі може перетворювати постановку задачі у складні фази, що, при виконанні, забезпечують розв’язок. У більш складних ситуаціях, проблемний процесор має бути здатний будувати моделі, потрібні для вирішення проблеми.

Інструментальний підхід: потреба у розробці інструментів програмного забезпечення для побудови СППР виникла внаслідок того, що представники комп’ютерних наук та економічної практики роблять більший акцент на методології розробки та реалізації СППР. Інструментальний підхід являє собою нове поняття класифікації СППР, що виникло внаслідок цього. Можна виділити три рівні СППР на основі деталей задач, що розв’язуються, та технологічних інструментів, що застосовуються процесі розробки системи [34]:

1) спеціалізовані СППР;
2) генератори СППР;
3) інструментарій СППР.

Індивідуальні користувачі або група користувачів являють собою цільову аудиторію для спеціалізованих СППР. Вони дають можливість єдиній СППР або команді розв’язувати конкретні задачі в певних обставинах.

Генератор СППР представляє собою набір пов’язаних інструментів програмного забезпечення, що дозволяє швидко та просто генерувати СППР під власні потреби. Ці інструменти включають пошук, обробку даних та їх введення, моделювання, тощо. Система керування інформацією, для прикладу, створюється з числа компонентів, таких як мова моделювання, генерація звітів, пошук інформації, і набір інструментів для статистичного та фінансового аналізу.

Наступні складові концептуальної основи генератора СППР відображають точку зору користувача:

1) керування користувацьким інтерфейсом;
2) керування представлення даних та результатів;
3) керування аналізом;
4) керування системою;
5) керування даними.

Також слід реалізовувати наступні базові типи інтерфейсів:

1) меню;
2) мова команд;
3) базова мова формату “питання-відповідь”.

Складне сприйняття користувачем задачі, що вони намагаються вирішити, повинно підтримуватися керуванням представлення. Ці представлення можуть бути відображені як командні процедури, графи або таблиці.

Обслуговування бази даних моделі являє собою те, що має на увазі керування аналізом. Множина інструкцій може бути представлена у вигляді підзадачі аналізу для керування даними в контексті математичного моделювання. База даних моделі повинна бути додана з використанням додаткових аналітичних інструментів, які надаються СУБД. Системний

симулятор для навчання користувачів та координації операцій користувачів гарантуються системним адміністратором.

Система управління базою даних або СУБД використовується для реалізації керування даними. Вона має мати в наявності здатність керування словником даних, що дає можливість створення інших слоників на основі цієї бази, так як словник моделей або візуальний словник.

Система REGIMES, котра акцентується на персональних комп'ютерах, може розглядатися як прототип генератора [34]. Командний процесор, діалоговий процесор, процесор відображення результатів, підсистема керування регресійним аналізом і три словники утворюють даний генератор.

Нові спеціалізовані мови, вдосконалені операційні системи, інструменти сумісного використанні інформації, проекції кольорових графічних зображень та інші складні інструменти доступні розробникам СППР завдяки інструментарію СППР. Внаслідок цього, вони можуть використовуватися для розробки генераторів СППР сумісно зі специфічними СППР.

4.1.2 Інтеграція системи підтримки прийняття рішень з актуарними моделями і вибір ознак

Для розробки СППР використовувалися актуарні дані страхової компанії. Датасет було взято з Singapore Actuarial Society. Кожен страховий поліс від нещасних випадків був використаний під час настання нещасного випадку.

Дані містили такі ознаки, перелічені нижче.

1. **DateOfTimeAccident**: категорійна змінна, що містить інформацію про дату нещасного випадку.
2. **DateReported**: категорійна змінна, що містить інформацію про дату сповіщення про нещасний випадок.
3. **Age**: неперервна змінна, що містить інформацію про вік клієнта. Для Байєсівської мережі дана величина була дискретизована за допомогою

використання наступних інтервалів: ‘Between 18 and 25’, ‘Between 25 and 50’ і ‘More than 50’. Вершина в мережі Байєса позначається як A.

4. **Gender**: категорійна змінна, що містить інформацію про стать клієнта. Містить два можливих значення: ‘M’ — чоловіча, ‘F’ — жіноча. Вершина в мережі Байєса позначається як G.

5. **MaritalStatus**: категорійна змінна, що містить інформацію про шлюбний статус клієнта. Включає наступні значення: ‘M’ — одружений, ‘S’ — неодружений, ‘U’ — невідомо. Вершина в мережі Байєса позначається як M.

6. **DependentChildren**: дискретна змінна, що містить інформацію щодо числа дітей страхового клієнта. Його діапазон становить від 1 до 2. Вершина в мережі Байєса позначається як CH.

7. **WeeklyWages**: неперервна змінна, що містить інформацію щодо щотижневої заробітної плати клієнта. Для Байєсівської мережі величина було дискретизована з використанням інтервалів: ‘Less than 100’, ‘Between 100 and 1000’ і ‘Between 100 and 50000’. Вершина в мережі Байєса позначається як W.

8. **PartTimeFullTime**: категорійна змінна, що містить інформацію про режим роботи. Включає наступні значення: ‘P’ — працює на півставки, ‘F’ — працює на повну ставку. Вершина в мережі Байєса позначається як PFT.

9. **HoursWorkedPerWeek**: неперервна змінна, що містить інформацію щодо кількості робочих годин клієнта.

10. **DaysWorkedPerWeek**: дискретна змінна, що містить інформацію про кількість робочих днів кожного страхового клієнта. Його діапазон становить від 1 до 7 днів. Вершина в мережі Байєса позначається як D.

11. **ClaimDescription**: категорійна змінна, що містить інформацію щодо опису страхового позову.

12. **InitialIncurredCalimsCost**: неперервна змінна, що містить інформацію про початковий розмір страхових виплат для клієнта.

13. **UltimateIncurredClaimedCost**: неперервна змінна, що містить інформацію щодо страхових виплат клієнта. Представляє собою цільову змінну для УЛМ та Байєсівської мережі. Для останньої моделі величина була

дискретизована з використанням наступних інтервалів: ‘Between 100 and 1000’, ‘Between 1000 and 50000’ і ‘Between 50000 and 100000’. Вершина в мережі Байєса позначається як CL.

При побудові УЛМ для всіх неперервних змінних застосовувалась нормалізація до одиничного інтервалу, а для категорійних змінних використовувався One-hot encoding. Також попередньо видалялися дані клієнтів, кількість робочих днів яких дорівнювала одиниці.

У цьому дослідженні дані зберігаються в текстовому файлі формату CSV.

Відбір змінних реалізовується за допомогою наступних етапів.

Крок 1. Спочатку предикторні ознаки обираються за допомогою тесту Хі-квадрат. З множини, що складалася з 12 ознак, тест обрав дев’ять ознак: Age, Gender, MaritalStatus, DependentChildren, DependentsOther, WeeklyWages, PartTimeFullTime, HoursWorkedPerWeek, DaysWorkedPerWeek, InitialIncurredCalimsCost. Усім ознакам, що пройшли тест нараховувалися по одному балу з метою обчислення загальної ваги кожної з ознак.

Крок 2. Далі ознаки проходять відбір за допомогою методу RFE. З того ж початкового набору алгоритм виділив сім предикторних змінних: Age, Gender, MaritalStatus, DependentChildren, WeeklyWages, PartTimeFullTime, DaysWorkedPerWeek.

Крок 3. Насамкінець предикторні ознаки обираються шляхом застосування методу випадкового лісу. Метод обрав з вхідної множини змінних 5 ознак: Age, Gender, DependentChildren, PartTimeFullTime, DaysWorkerPerWeek.

Насамкінець потрібно обчислити остаточну оцінку кожної ознаки, і на основі неї приймається рішення щодо подальшого використання при побудові моделей для аналізу актуарних ризиків. Остаточний результат наведено в таблиці 4.1.

Ті ознаки, що набрали сумарно два бали і більше, використовувалися при подальшій побудові моделей і розробці СППР: Age, Gender, MaritalStatus,

DependentChidlren, WeeklyWages, PartTimeFullTime, DaysWorkedPerWeek. Інші ознаки вважаються незначущими.

Таблиця 4.1 — Результати відбору ознак за тестом Хі-квадрат, RFE та методом випадкового лісу

Предикторна ознака	*Тест Хі-квадрат*	*Метод RFE*	*Метод випадкового лісу*	*Остаточна оцінка*
DateTimeOfAccident	0	0	0	0
DateReported	0	0	0	0
Age	1	1	1	3
Gender	1	1	1	3
MaritalStatus	1	1	0	2
DependentChidlren	1	1	1	3
WeeklyWages	1	1	0	2
PartTimeFullTime	1	1	1	3
HoursWorkedPerWeek	1	0	0	1
DaysWorkedPerWeek	1	1	1	3
ClaimDescription	0	0	0	0
InitialIncurredCalimsCost	1	0	0	1

4.1.3 Використання сучасних технологій

Для розробки СППР використовувалася мова програмування Python версії 3.12. Для створення інтерфейсу програмного забезпечення використовувався Qt Designer.

Під час створення програми використовувалися такі пакети:

— PyQt6: реалізація функціонування інтерфейсу;

— Pandas: робота з даними;

— Scikit-Learn: попередня обробка даних (One-Hot encoding);

— Bambi: побудова УЛМ методом Монте-Карло;

— Pgmpy: побудова мережі Байєса;

— Numpy: робота з масивами і використання елементів лінійної алгебри;

— Arviz, Matplotlib, Seaborn: візуалізація даних.

4.2 Приклади застосування системи підтримки прийняття рішень

Процес роботи СППР можна представити у вигляді схеми, наведеної на рисунку 4.5.

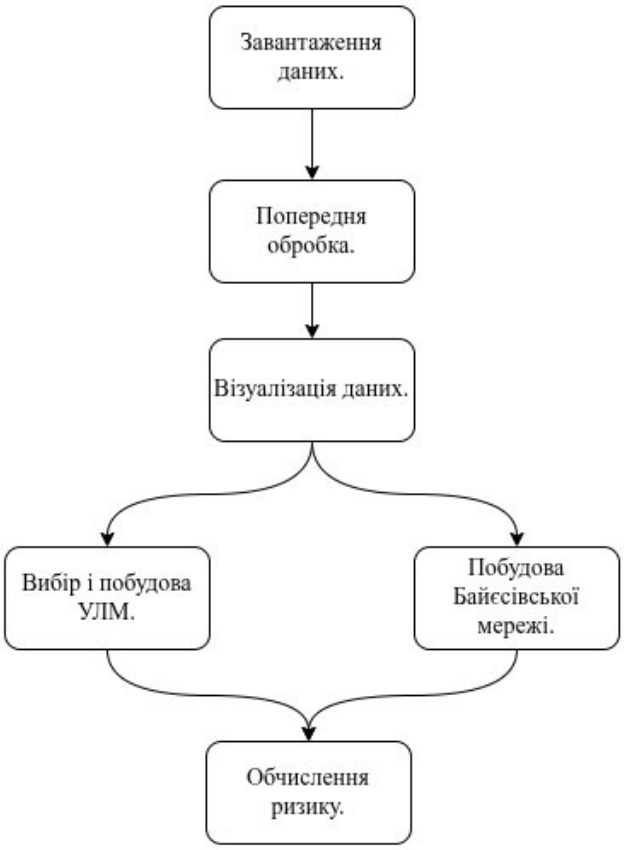

Рисунок 4.5 — Схема функціонування СППР

У наступних підрозділах наведено детальний опис кожного етапу роботи.

4.2.1 Детальний опис інтерфейсу та кроків користувача

Інтерфейс розробленої СППР показано рисунку 4.6.

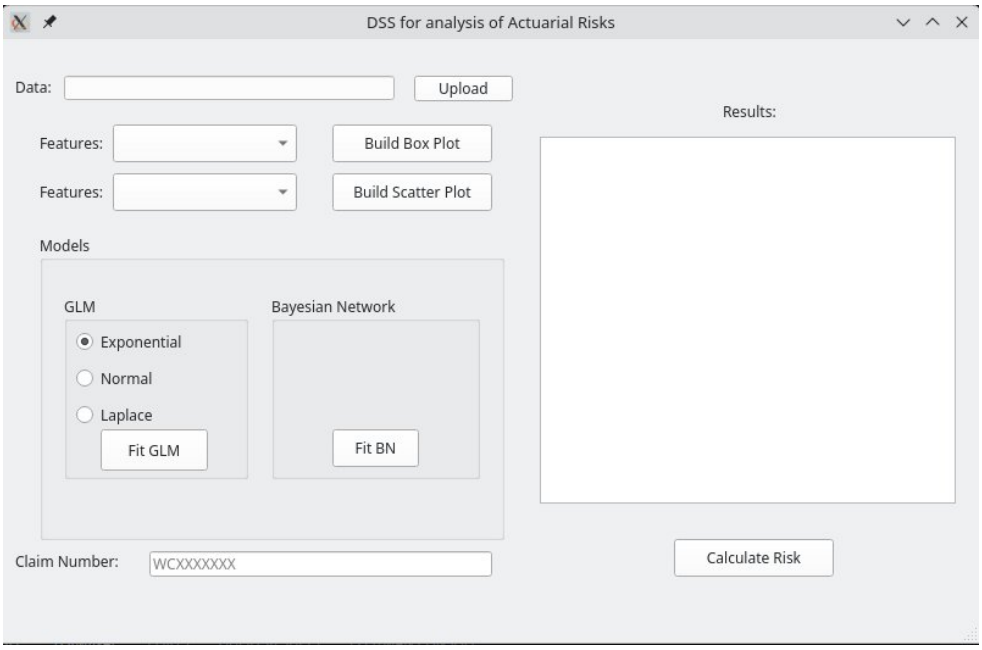

Рисунок 4.6 — Користувацький інтерфейс СППР

Після запуску програми користувач має завантажити файл для обчислення ризику. Для цього він має натиснути на кнопку Upload і перед ним з’явиться діалогове вікно для вибору файлу формату CSV. Приклад вікна наведено на рисунку 4.7.

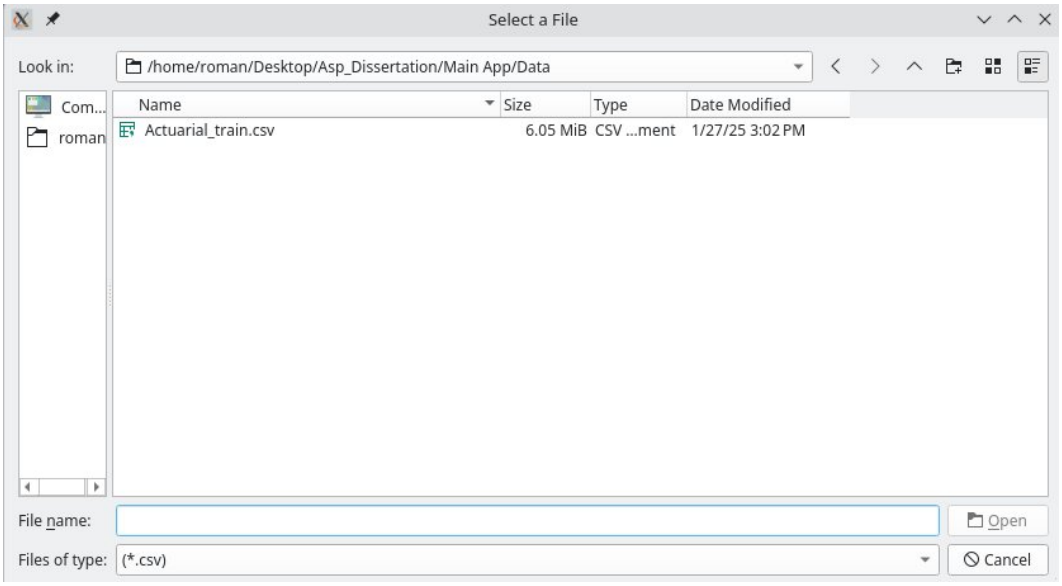

Рисунок 4.7 — Діалогове вікно вибору

Після успішного завантаження даних створюються копії даних для УЛМ та Байєсівської мережі і для двох останніх відбувається попередня обробка даних, що включає в себе нормалізацію та використання One-hot encoding для УЛМ та дискретизації неперервних даних для мережі Байєса. Після цього заповнюються випадаючі списки після мітки Features. Перший містить усі назви категорійних та дискретних полів, а другий — назви полів неперервних даних. Обираючи елементи та натискаючи на кнопку Build Box Plot або Build Scatter Plot користувач може візуалізувати дані страхових показників по відношенню до цільової змінної у вигляді діаграми розмаху або діаграми розсіювання.

Після цього користувач має обрати за своїм бажанням одну із УЛМ та натиснути на кнопку Fit GLM для побудови УЛМ, що займе певний час. Після успішної побудови СППР сповістить користувача про успішне завершення процедури. Екран сповіщення про успішну побудову УЛМ наведено на рисунку 4.8.

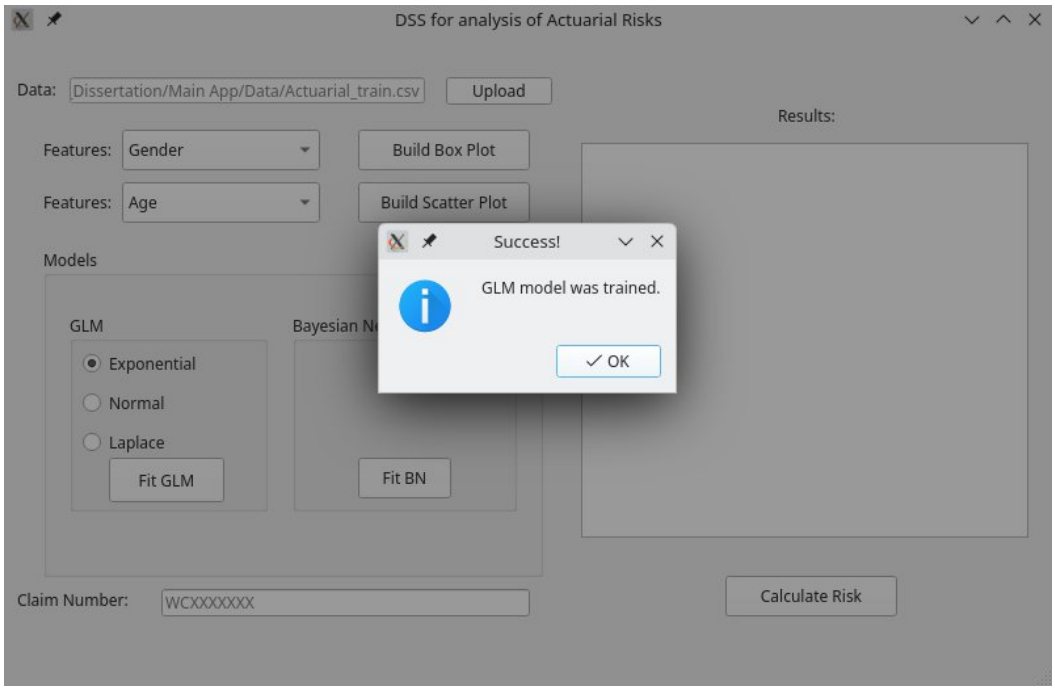

Рисунок 4.8 — Вікно сповіщення про успішну побудову УЛМ

Після цього програма відобразить графік результату навчання УЛМ у вигляді кривої розподілу цільової змінної та розподілу створеної УЛМ. Також СППР відобразить на головному екрані результати налаштування у вигляді значень метрик якості побудови моделі.

Приклад результатів наведено на рисунку 4.9.

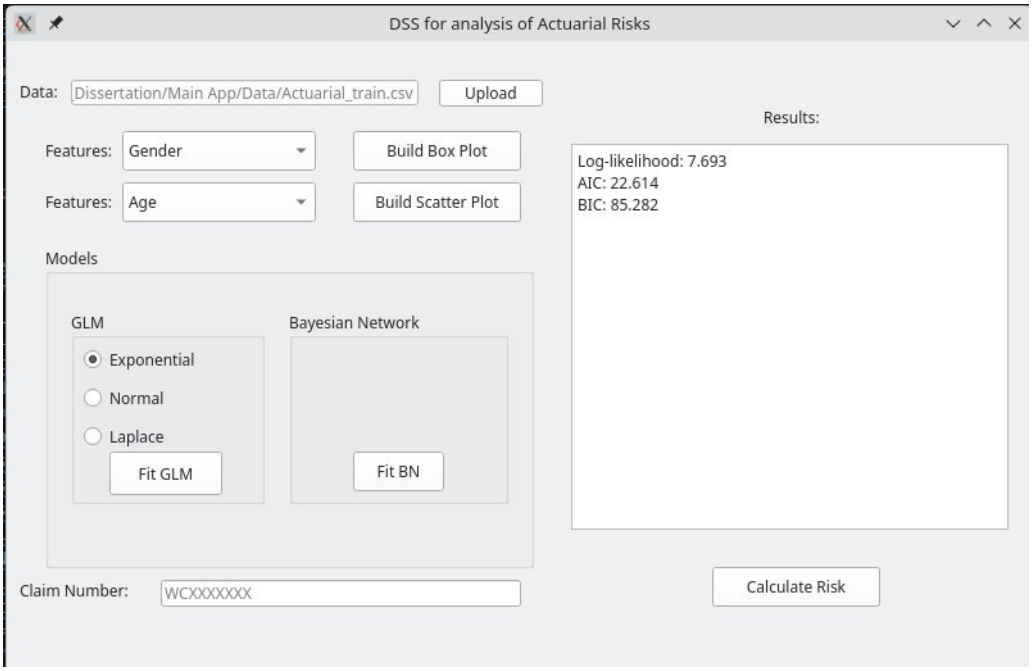

Рисунок 4.9 — Результат навчання УЛМ в форматі значень метрик якості моделі

Визначившись з типом УЛМ, користувач також має побудувати Байєсівську мережу. Для цього він має натиснути на кнопку Fit BN. У разі успішного завершення процедури, програма сповістить у вигляді інформаційного вікна, приклад якого наведено на рисунку 4.10.

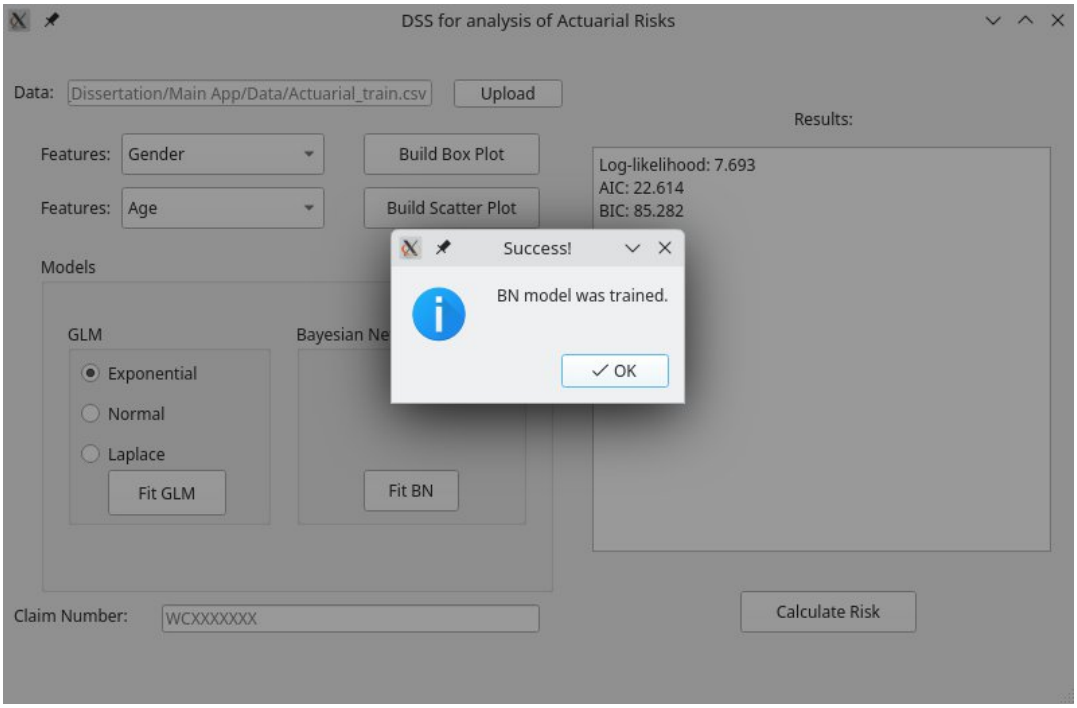

Рисунок 4.10 — Вікно сповіщення про успішну побудову Байєсівської мережі

Після того, як УЛМ та мережа Байєса були побудовані, користувач може обчислити ризик наступним чином. Для цього він має ввести в поле номер страхового позову формату WCXXXXXXX, де X являє собою число від 0 до 9. Після того як користувач натисне на кнопку Calculate Risk, СППР почне пошук даних. Якщо дані було знайдено, програма надасть інформацію про клієнта, втрати в грошових одиницях, обчислених УЛМ та ймовірності грошових втрат в інтервалах, обчислених мережею Байєса. Приклад результату розрахунку ризику наведено на рисунку 4.11.

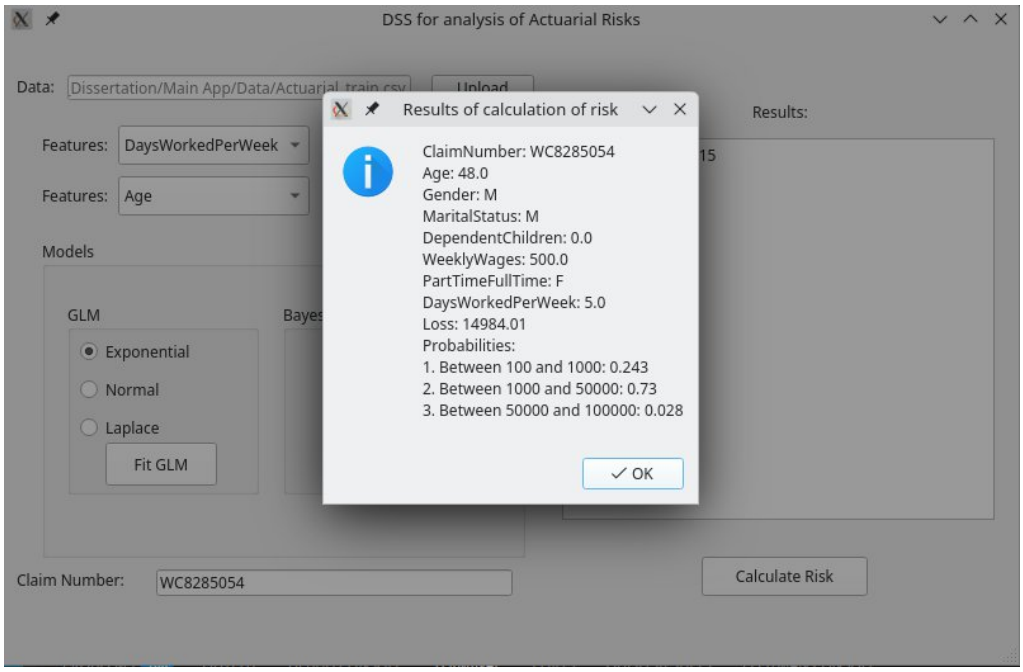

Рисунок 4.11 — Приклад результату обчислення ризику

Як можна побачити, побудована СППР здатна розрахувати ризик втрат у грошових одиницях та його ймовірності на основі повної інформації щодо страхового клієнта.

4.2.2 Побудова графіків

Як було зазначено вище СППР дозволяє користувачу візуалізовувати дані страхових показників по відношенню до цільової змінної у вигляді діаграм розмаху для дискретних та категорійних величин так і діаграм розсіювання для неперервних величин.

Діаграми розмаху СППР наведено на рисунку 4.12.

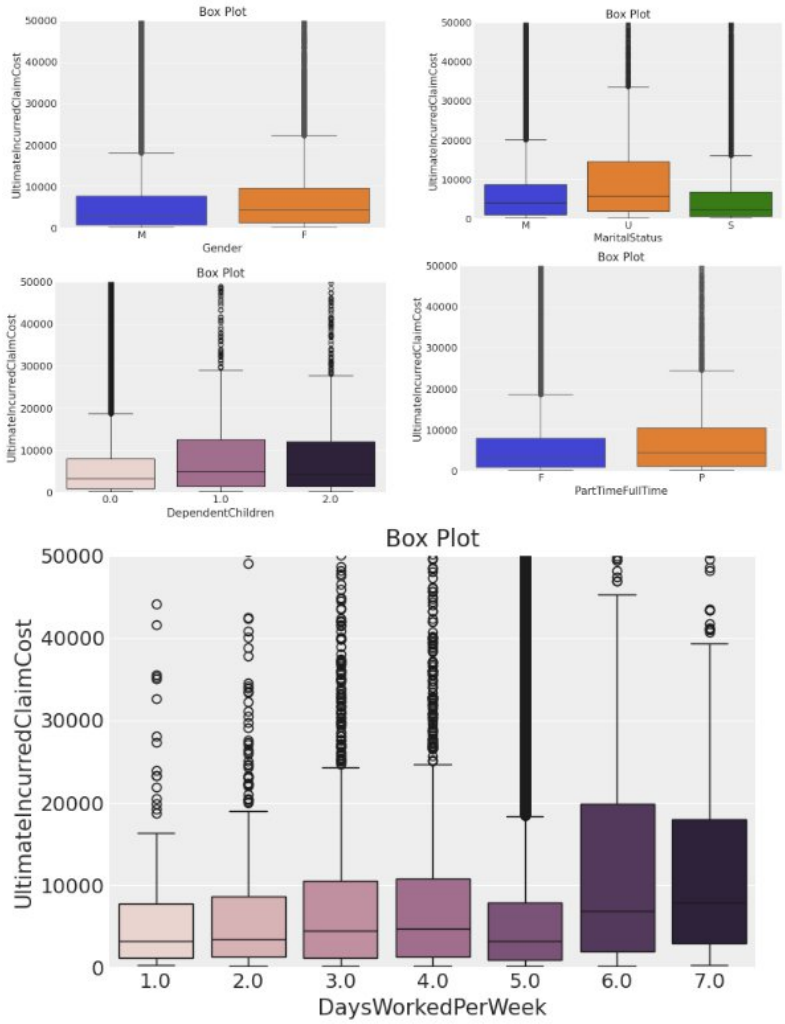

Рисунок 4.12 — Графіки розмаху для показників Gender, MaritalStatus, DependentChildren, PartTimeFullTime та DaysWorkedPerWeek

Діаграми розсіювання наведено на рисунку 4.13.

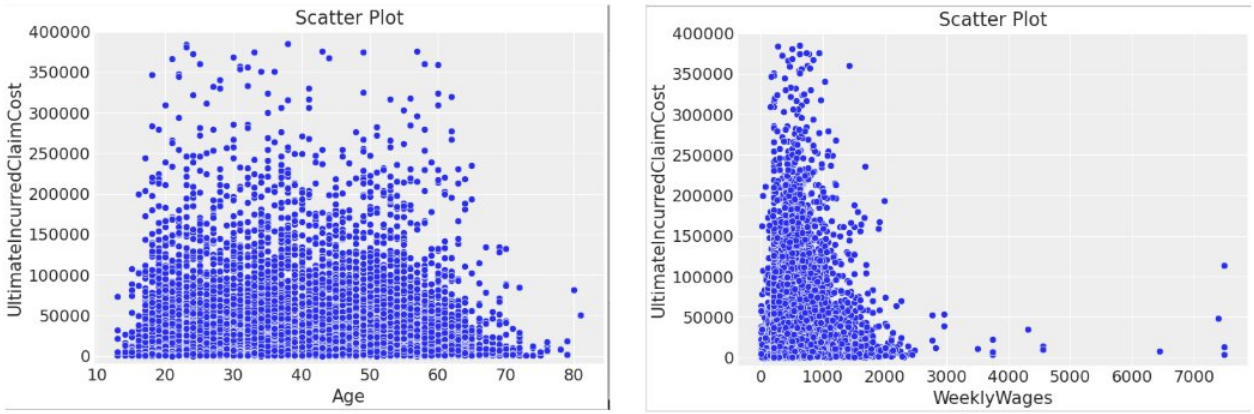

Рисунок 4.13 — Графіки розсіювання для показників Age та WeeklyWages

4.2.3 Можливість вибору побудови моделей

СППР дає можливість користувачу обрати наступні види УЛМ:

1) нормальний розподіл з логарифмічною функцією зв'язку;
2) експоненціальний розподіл з логарифмічною функцією зв'язку;
3) розподіл Лапласа з тотожною функцією зв'язку.

4.2.4 Побудова мережі Байєса в системі підтримки прийняття рішень

Топологічна структура мережі Байєса наведена на рисунку 4.14.

Для побудови таблиць умовних ймовірностей застосовував підхід емпіричного розподілу для Байєсівських мереж [70]. Вектори $d_1,...,d_M$ задають датасет D для змінних X; кожен d_i розглядається як випадок і вказує на певний зразок змінних X. Датасет вважається повним у даному дослідженні.

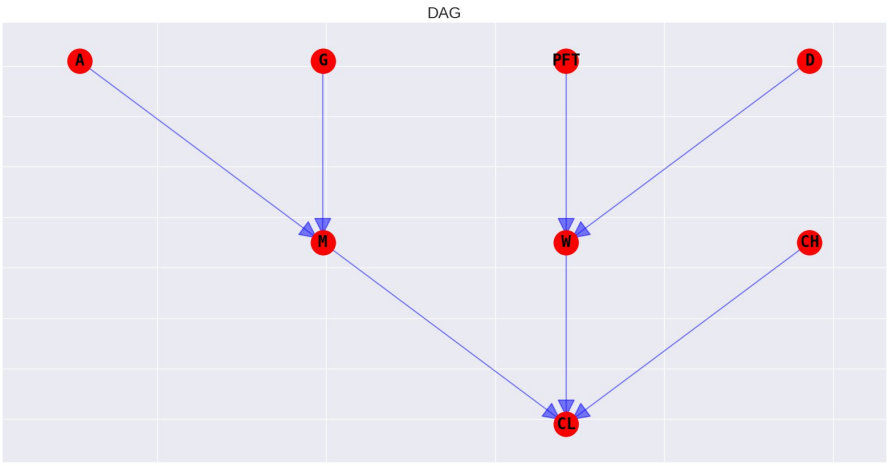

Рисунок 4.14 — Топологічна структура Байєсівської мережі для СППР

Визначення емпіричного розподілу для повного датасету має такий вираз:

$$P_D(k) = \frac{Num(k)}{M}, \tag{4.1}$$

де $Num(k)$ — кількість випадків d_i в датасеті D, що містить подію k.

Результати моралізації та перетворення моралізованого графу в неорієнтований наведено на рисунках 4.15 та 4.16.

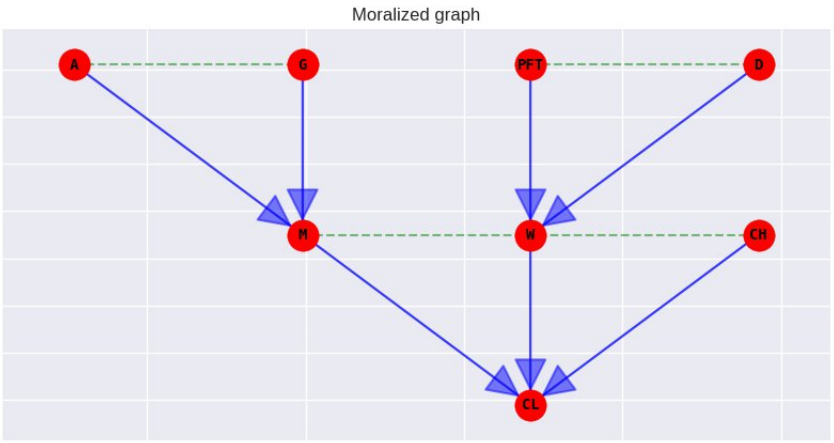

Рисунок 4.15 — Моралізований граф

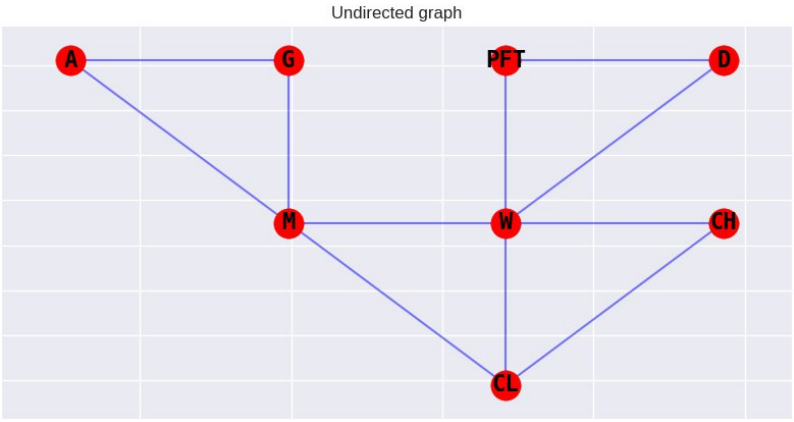

Рисунок 4.16 — Неорієнтований граф

Далі мережа Байєса була перетворена в дерево об’єднань, структура якої подана на рисунку 4.17.

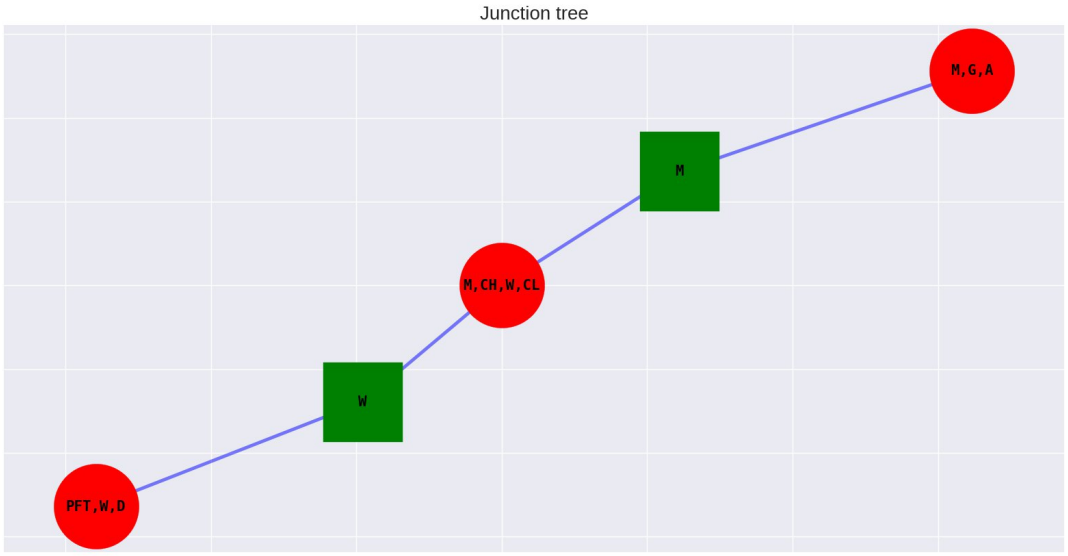

Рисунок 4.17 — Дерево об’єднань

Кліка “M, CH, W, CL” виконує функцію кореня, а кліки “M, G, A” та “PFT, W, D” — листків.

Логістичну регресію було застосовано для порівняння результатів роботи мереж Байєса після введення даних у вузли. Алгоритм класифікації, відомий як

логістична регресія, прогнозує бінарно-залежний результат за рахунок застосування декількох незалежних факторів. Це дуже ефективний метод для розуміння того, як інформація, вказівки або певна подія пов’язані один з одним [71].

Застосовуючи множину вхідних змінних, логістична регресія має на меті моделювати ймовірність певного результату. Вихідна змінна в логістичній регресії є бінарною, що означає, що вона може мати лише одне з двох можливих значень. Будь-який фактор, що впливає на цільову залежну змінну є незалежною змінною. Разом з тим, вхідні змінні можуть бути категорійними, неперервними або їх комбінацією.

Результат перетворюється в категорійне значення за допомогою використання функції сигмоїди. Сигмоїдальна функція, що задає S-подібну криву та генерує ймовірнісне значення між 0 та 1, використовується в машинному навчанні для перетворення прогнозів у ймовірності.

Модель можна записати таким чином:

$$P(X) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}. \quad (4.2)$$

Цей вираз можна перезаписати таким чином:

$$P(X) = \frac{1}{1 + e^{-X}}, \quad (4.3)$$

де X — лінійна комбінація пояснювальних змінних,

$P(X)$ — ймовірність настання певної події.

Поліноміальна логістична регресія використовується, коли залежна змінна містить більше, ніж дві категорії, але не містить впорядкованих підкатегорій. Інакше кажучи, немає внутрішнього порядку, а категорії є взаємовиключними.

Результати ймовірнісних висновків вершини CL Байєсівської мережі і логістичної регресії після запитів наведено в таблицях 4.2-4.5.

Таблиця 4.2 — Ймовірнісний висновок вершини CL після запиту W = "*Between* 1000 *and* 50000"; $M = S$; $CH = 0$

Стан	*Ймовірність (Байєсівська мережа)*	*Ймовірність (логістична регресія)*
Between 100 and 1000	0.023	0.106
Between 1000 and 50000	0.895	0.828
Between 50000 and 100000	0.082	0.066

Таблиця 4.3 — Ймовірнісний висновок вершини CL після запиту W = "*Less than* 100"; $M = S$; $CH = 1$

Стан	*Ймовірність (Байєсівська мережа)*	*Ймовірність (логістична регресія)*
Between 100 and 1000	0.368	0.393
Between 1000 and 50000	0.621	0.603
Between 50000 and 100000	0.011	0.004

Таблиця 4.4 — Ймовірнісний висновок вершини CL після запиту W = "*Less than* 100"; A = "*Between* 18 *and* 25"; $D = 5$

Стан	*Ймовірність (Байєсівська мережа)*	*Ймовірність (логістична регресія)*
Between 100 and 1000	0.025	0.095
Between 1000 and 50000	0.965	0.698
Between 50000 and 100000	0.000	0.007

Таблиця 4.5 — Ймовірнісний висновок вершини CL після запиту $G = M; A =$ "*More than* 50

Стан	*Ймовірність (Байєсівська мережа)*	*Ймовірність (логістична регресія)*
Between 100 and 1000	0.243	0.081
Between 1000 and 50000	0.727	0.908
Between 50000 and 100000	0.030	0.011

З таблиць вище можна зробити висновок, що ймовірності мереж Байєса часто показують набагато кращий результат, ніж ймовірності логістичної регресії.

4.2.5 Аналіз результатів моделювання з використанням системи підтримки прийняття рішень

Для оцінювання якості побудованих УЛМ використовувався логарифм максимальної правдоподібності, інформаційний критерій Акайке і Байєсівський інформаційний критерій.

Отримані значення метрик для трьох моделей подано в таблиці 4.6.

Таблиця 4.6 — Значення метрик якості моделей для УЛМ

УЛМ	*Log-Likelihood*	*AIC*	*BIC*
Exponential	7.573	22.854	85.522
Normal	1.851	34.297	96.965
Laplace	3.605	30.79	93.458

Графік результатів побудови УЛМ наведено на рисунках 4.18, 4.19 та 4.20.

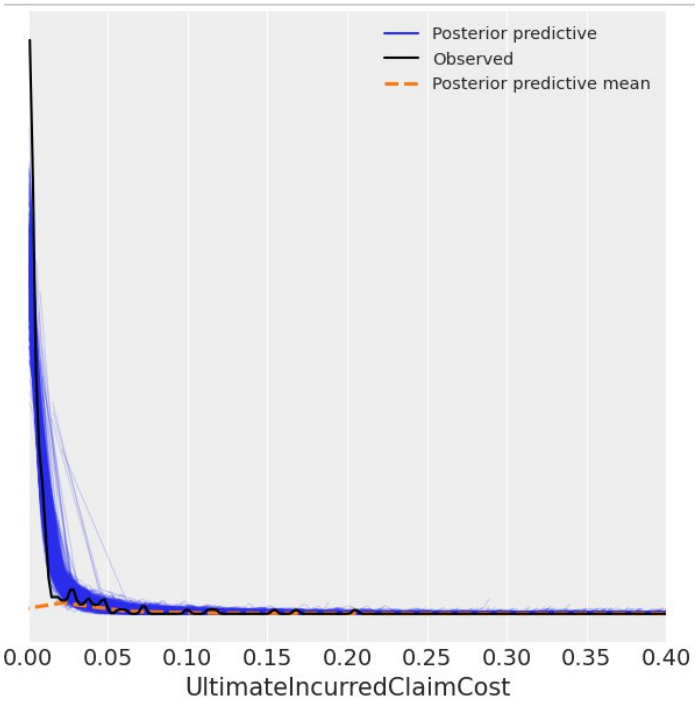

Рисунок 4.18 — Результат побудови експоненційної УЛМ

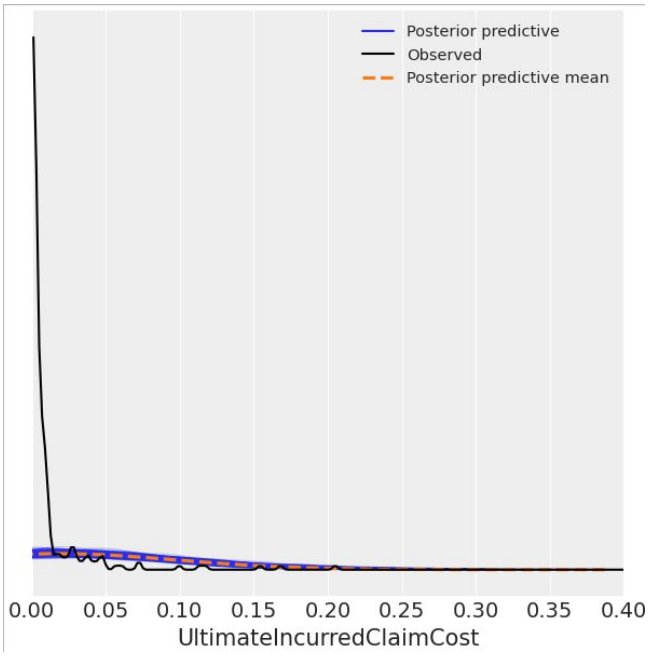

Рисунок 4.19 — Результат побудови гаусівської УЛМ

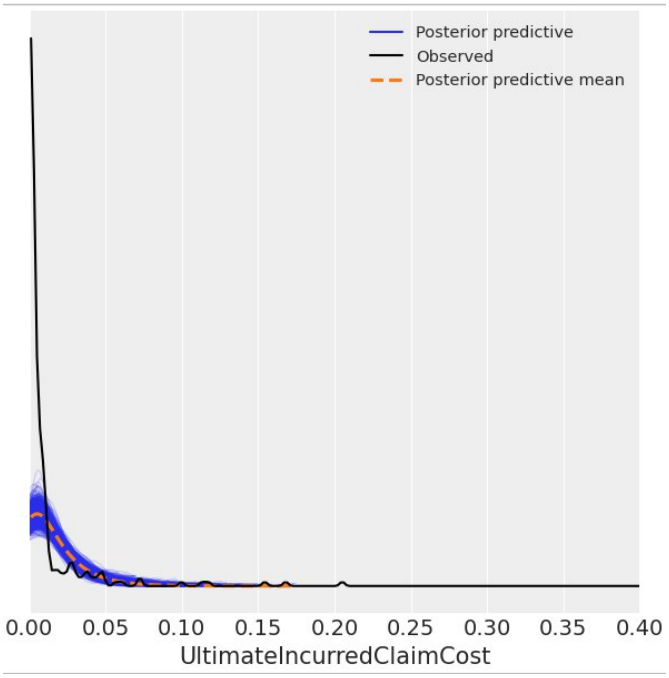

Рисунок 4.20 — Результат побудови УЛМ з розподілом Лапласа

На основі таблиці 4.5 та рисунків 4.18-4.20 можна зробити висновок, що експоненціальна УЛМ показала найкращі результати.

4.2.6 Аналіз рівня ризику на основі отриманих даних

Розглянемо приклади обчислення значення ризику за допомогою СППР після побудови обраної УЛМ та мережі Байєса. У подальшому використовується експоненційна УЛМ.

Розглянемо випадок страхового позову з номером WC6683299. Результати обчислення подано на рисунку 4.21.

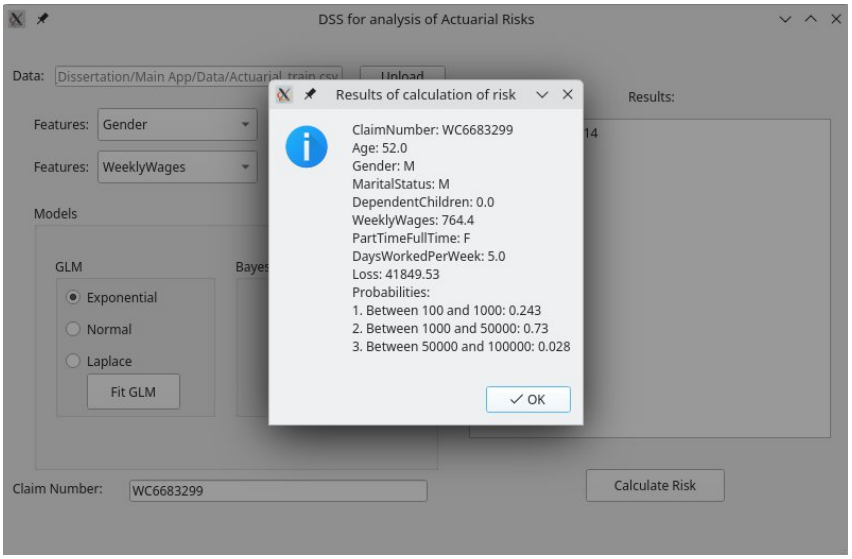

Рисунок 4.21 — Результати обчислення СППР для страхового позову WC6683299

За датасетом виплати згідно позову становили 40519.72, що свідчить про те, що УЛМ обчислило втрати, тобто виплати, доволі точно.

Розглянемо інший випадок зі страховим позовом, номер якого WC7639313. Результати обчислення наведено на рисунку 4.22.

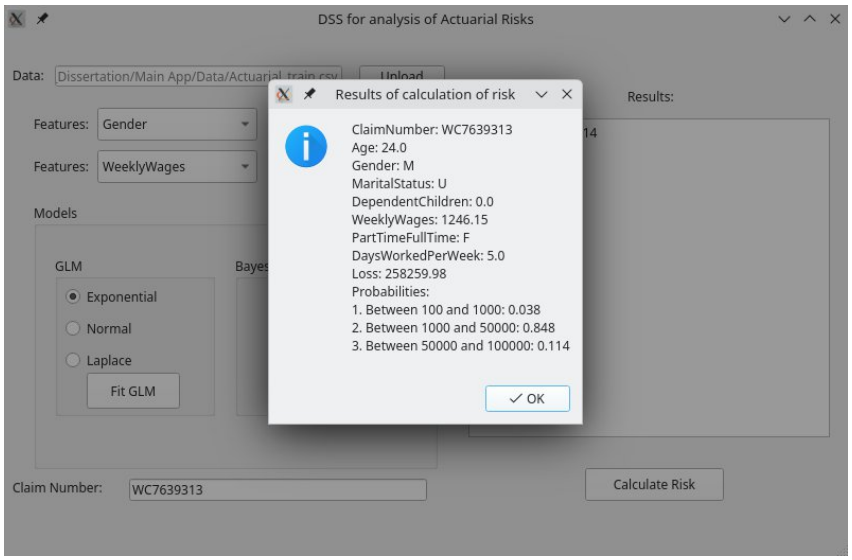

Рисунок 4.22 — Результати обчислення СППР для страхового позову WC7639313

За результатами можна помітити, що ймовірності для трьох інтервалів суттєво змінились, хоч обчислені втрати більші за виплати в датасеті, що становить 59868.41.

Наостанок обчислимо ризик для страхового позову з кодом WC9366329. Результати обчислення наведено на рисунку 4.23.

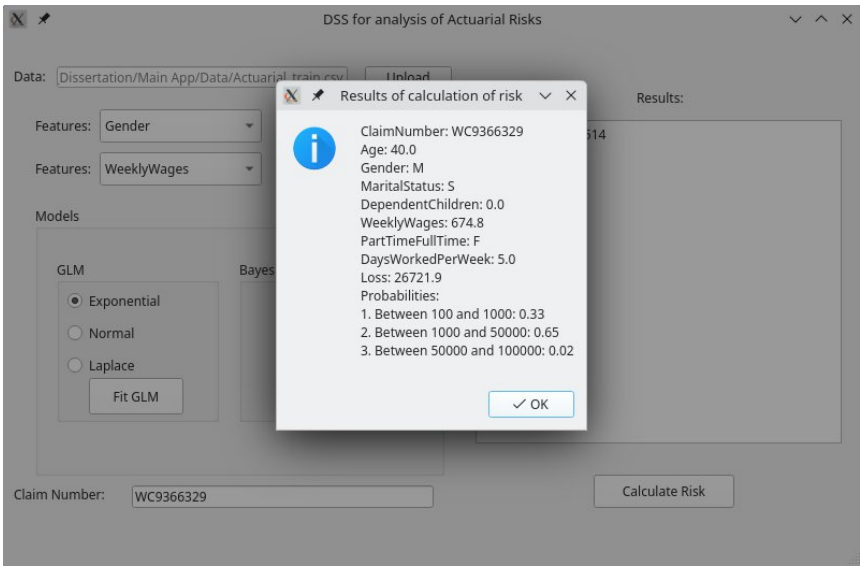

Рисунок 4.23 — Результати обчислення за допомогою СППР для страхового позову WC9366329

Можна помітити, що ймовірності також змінилися. Крім того, обчислені втрати вище згідно даних, що становлять 2310,69.

4.2.7 Масштабування роботи системи підтримки прийняття рішень

Масштабування роботи СППР розглядалося з таких сторін.

Масштабування функцій розподілу УЛМ. У даному дослідженні розглядаються три варіанти функцій розподілу, а саме: експоненційна, нормальна і з розподілом Лапласа, що можуть бути використані для побудови УЛМ. За рахунок цього користувач може обрати один з підходів або одночасно зробити експерименти з трьома підходами і порівняти отримані результати.

Тривалість побудови трьох зазначених УЛМ подано в таблиці 4.7.

Таблиця 4.7 — Тривалість побудови УЛМ

УЛМ	*Тривалість*
Exponential	14 хвилин, 18 секунд.
Normal	12 секунд.
Laplace	15 хвилин.

Можна побачити, що час навчання експоненційної УЛМ є граничним для побудови УЛМ з задовільними значеннями метрик.

Масштабування з кількістю даних. Для отримання фінальних результатів ми спираємось на обсяги даних, які можуть бути опрацьовані в системі. Система може працювати як і з малою кількістю даних так і з десятками тисяч і з тисячами. Звичайно на цей результат будуть впливати час опрацювання і потужності.

Розглянемо випадок повної вибірки для експоненційного розподілу. Значення метрик якості побудови УЛМ подано в таблиці 4.8.

Таблиця 4.8 — Значення метрик якості експоненційної УЛМ з повною вибіркою

Метрика	*Значення*
Log-likelihood	7,596
AIC	22,807
BIC	160,493

Графік результату побудови УЛМ для випадку повної вибірки наведено на рисунку 4.24.

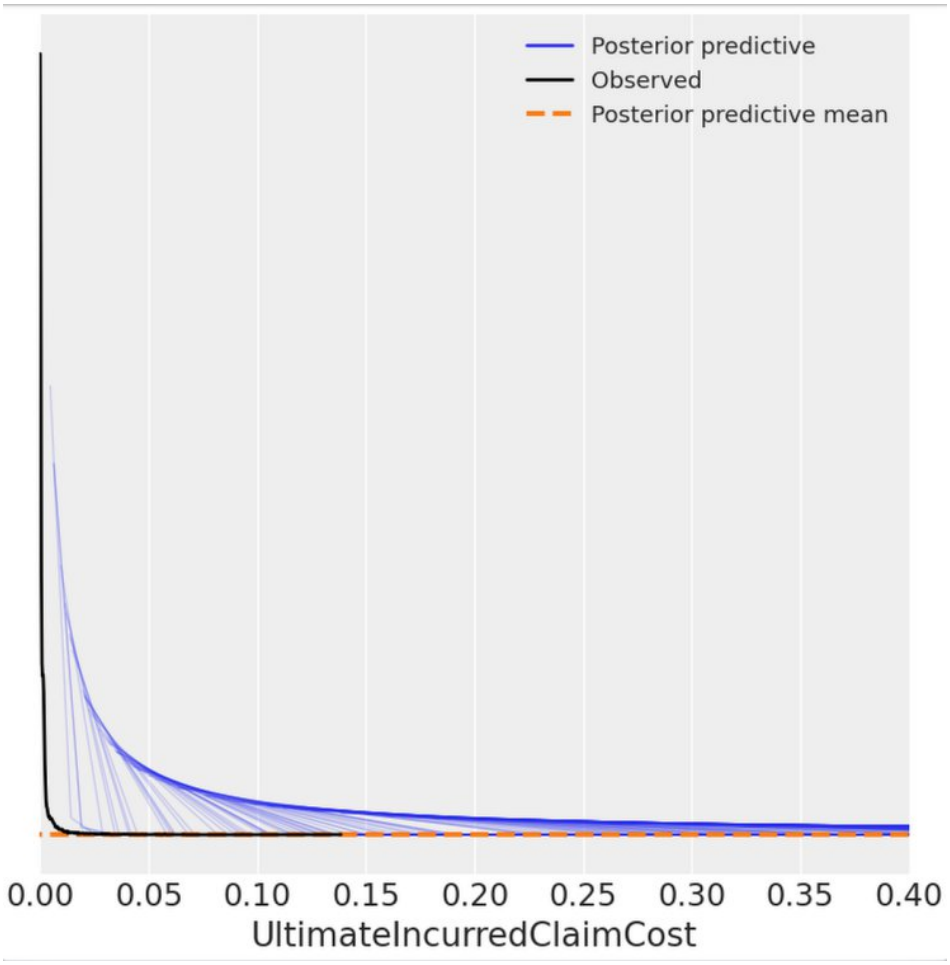

Рисунок 4.24 — Результат побудови експоненційної УЛМ для випадку повної вибірки

Результат обчислення ризику наведено на рисунку 4.25.

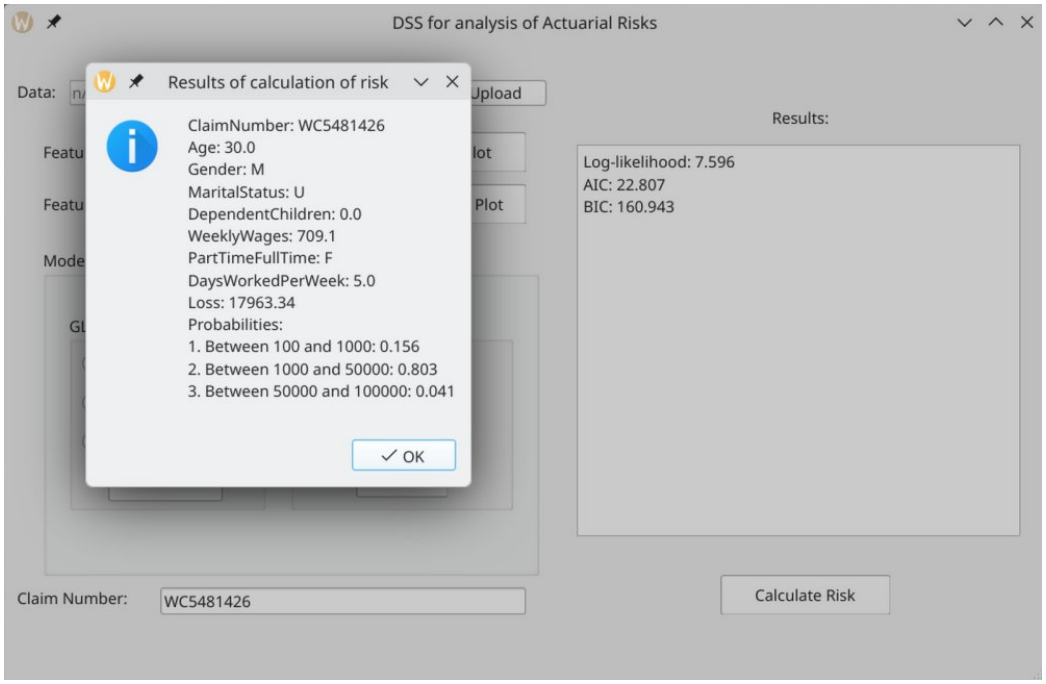

Рисунок 4.25 — Результат обчислення ризику

Розглянемо випадок вибірки з 900 записів. Значення метрик якості побудови УЛМ подано в таблиці 4.8.

Таблиця 4.9 — Значення метрик якості експоненційної УЛМ з вибіркою з 900 записів

Метрика	*Значення*
Log-likelihood	7,275
AIC	23,451
BIC	84,117

Графік результату побудови УЛМ для випадку повної вибірки наведено на рисунку 4.26.

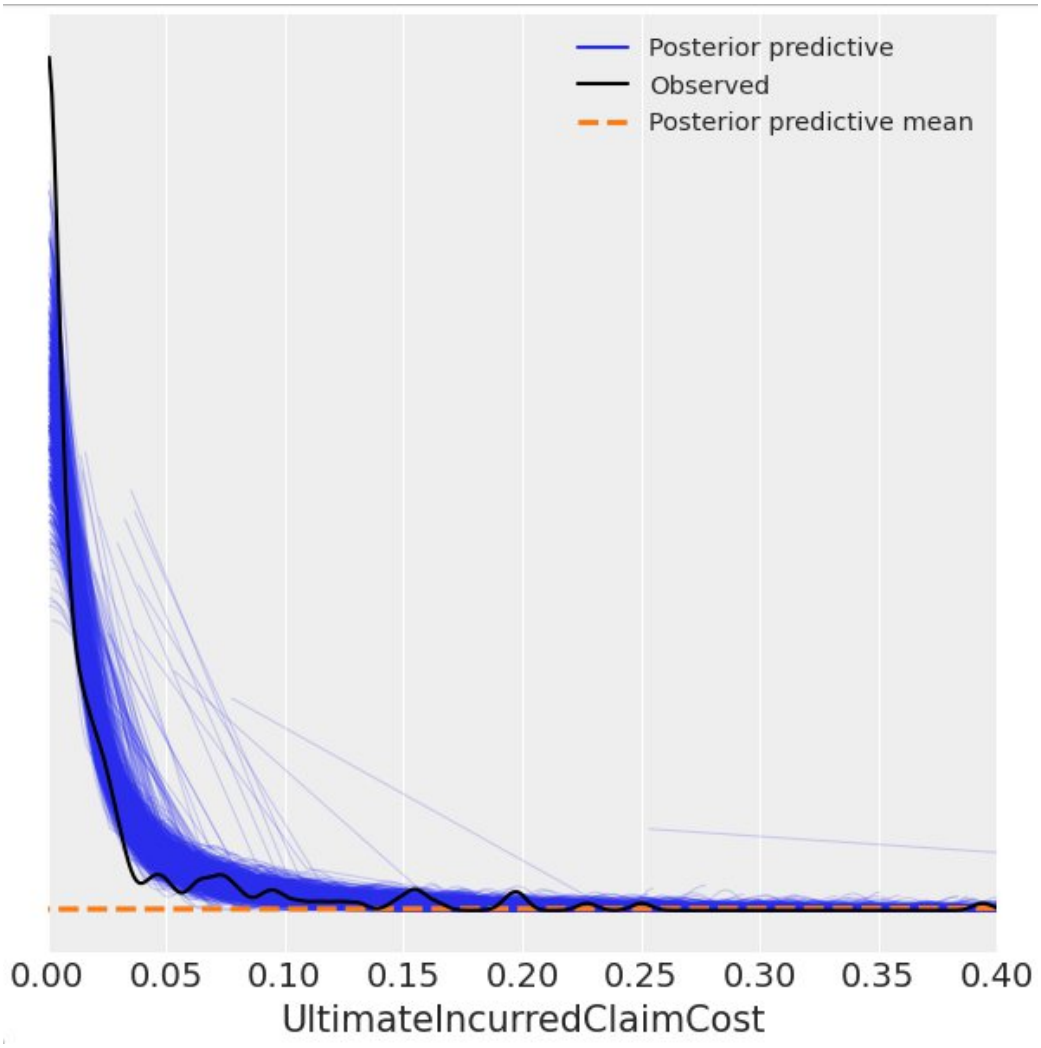

Рисунок 4.26 — Результат побудови експоненційної УЛМ для випадку вибірки з 900 записів

Результат обчислення ризику наведено на рисунку 4.27.

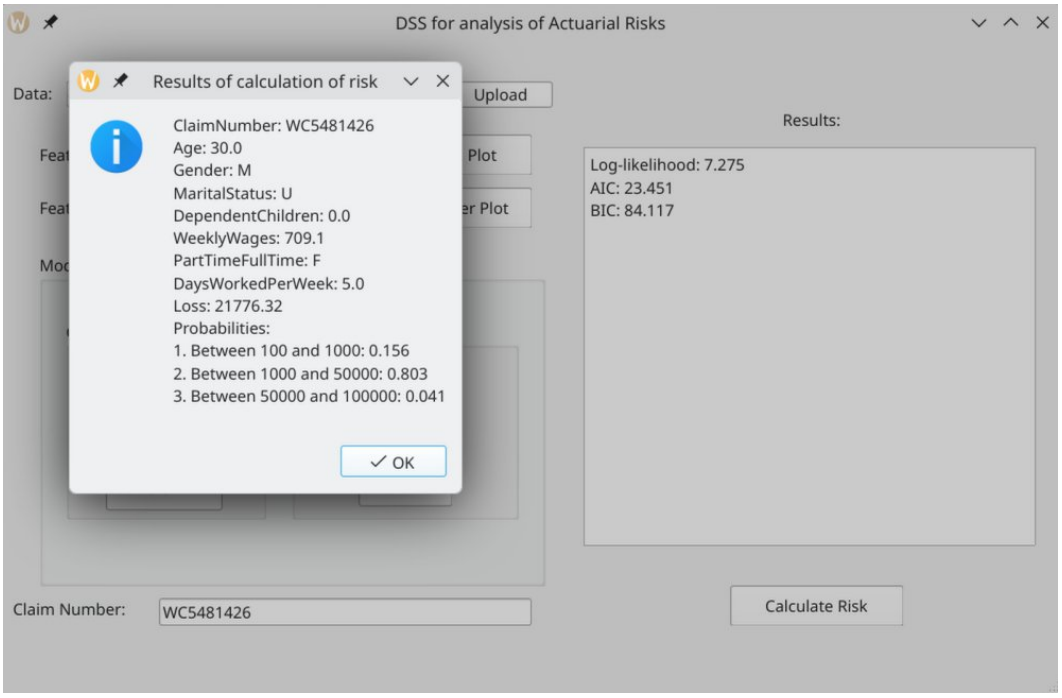

Рисунок 4.27 — Результат обчислення ризику

Насамкінець розглянемо випадок вибірки, що складається з 700 записів. Значення метрик якості побудови УЛМ подано в таблиці 4.8.

Таблиця 4.10 — Значення метрик якості експоненційної УЛМ з вибіркою з 700 записів

Метрика	*Значення*
Log-likelihood	5,505
AIC	26,99
BIC	82,881

Графік результату побудови УЛМ для випадку вибірки з 700 записів наведено на рисунку 4.28.

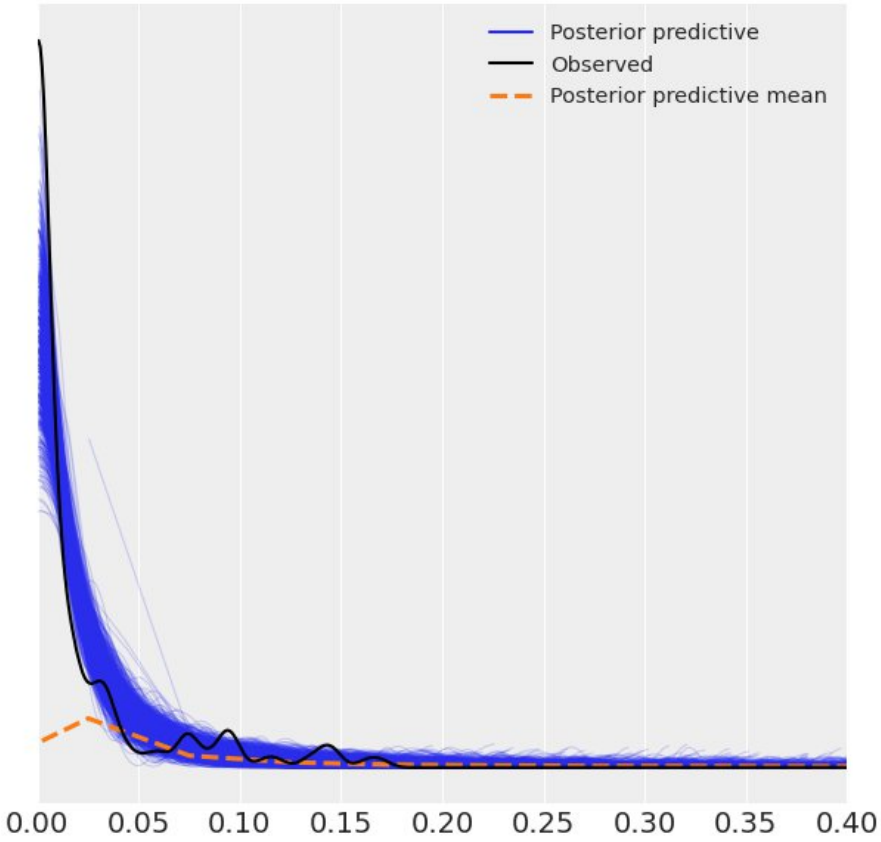

Рисунок 4.28 — Результат побудови експоненційної УЛМ для випадку вибірки з 700 записів

Результат обчислення ризику наведено на рисунку 4.29.

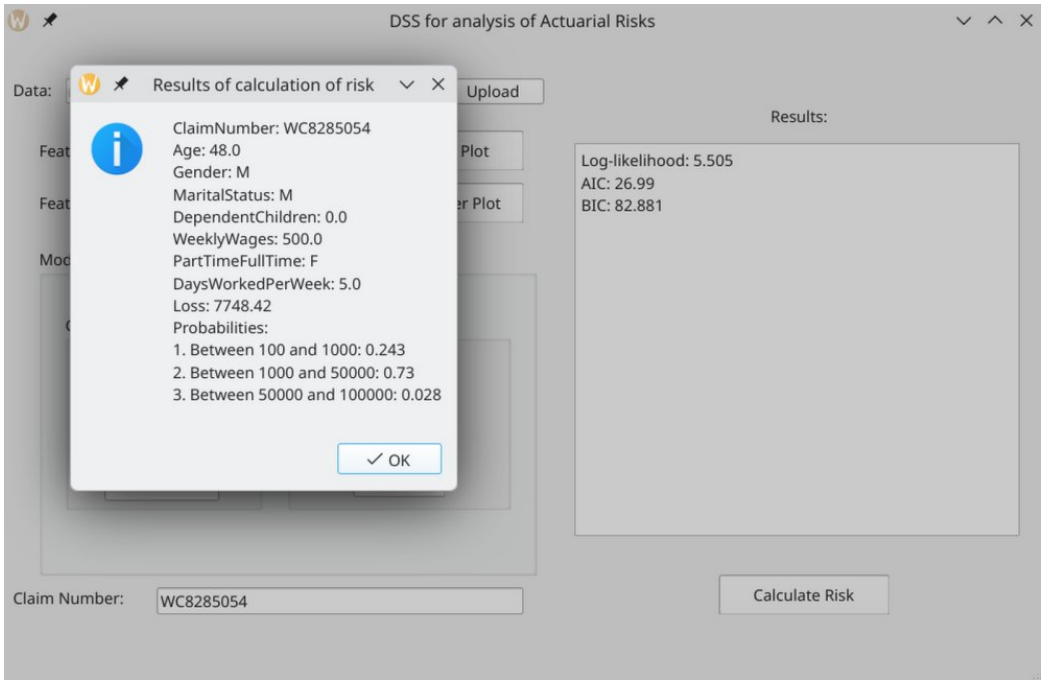

Рисунок 4.29 — Результат обчислення ризику для випадку вибірки з 700 записів

За трьома прикладами можна побачити, що зі зменшенням обсягу вибірки розподіли УЛМ за методом Монте-Карло для марківських ланцюгів наближуються до розподілу цільової змінної. Водночас логарифм максимальної правдоподібності зменшується, а критерій Акайке поступово зростає. Також варто звернути увагу, що при розмірності вибірки з 700 записів результат УЛМ був більш близьким до реальних даних.

4.3 Перспективи розвиту системи підтримки прийняття рішень у страхуванні

У подальших дослідженнях пропонується створити СППР у вигляді веб-додатку з використанням підходу REST API з можливістю додавання, оновлення або видалення даних про страхових клієнтів. Разом з цим розглядається можливість використання хмарних платформ.

В якості провайдерів хмарних сервісів можна розглянути представлені нижче [72].

Amazon Web Service (AWS). Пропонуючи набір сервісів для задоволення різноманітних потреб, що включає в собі зберігання даних, керування базами даних, аналітику, мережі, інтернет речей, мобільні розрахунки та бізнес рішення, AWS став лідером у сфері хмарних обчислень з моменту його заснування у 2006 році. Ці сервіси допомагають підприємствам у масштабуванні власної діяльності, мінімізації витрат та в прискоренні зростання. AWS представив себе як провідну хмарну платформу глобально за рахунок використання власного досвіду та стійкої інфраструктури. Дане глобальне охоплення, що розширюється, гарантує резервування та надійність, даючи можливість компаніям безпечно розгортати власні сервіси та додатки. Суміжно з широким діапазоном послуг та світового охоплення, AWS широко відома завдяки сильними протоколами безпеки та узгодженості, що є необхідним для бізнесу. Вона пропонує розширену архітектуру безпеки, керування ідентифікацією та доступом (IAM), шифрування даних та відповідає міжнародним стандартам сертифікації. За рахунок акценту на безпеку підприємства заохочуються оброблювати конфіденційні дані з більшою впевненістю з дотримування строгих нормативних стандартів.

Google Cloud Platform (GCP). У 2011 році Google зробив крок в області хмарних обчислень з GCP, що надає діапазон послуг для задоволення різноманітних вимог клієнтів. Запуск GCP був переломним моментом у бажанні Google у забезпеченні власній користувацькій базі новітніми технологічними пропозиціями. Підприємства можуть просто розгортати власні додатки при використанні потужної інфраструктури Google завдяки розширеній присутності мережевої зони, що гарантує відмінну доступність та надійність. GCP є серйозним конкурентом на ринку хмарних обчислень завдяки його відданості пропонувати безпечну, масштабовану та гнучку хмарну платформу, що відповідає змінним вимогам компаній в різних секторах. За рахунок Vertex AI, що модернізує розробку та використання моделей машинного навчання, GCP досяг значних просувань в області ШІ та машинного навчання. Також він надає BigQuery, безсерверне і масштабоване сховище даних, розроблене для швидких

SQL запитів та детального аналізу даних. Разом з тим, платформа Anthos від GCP пропонує єдине гібридне та мультихмарне рішення у сфері керування, що дозволяє підприємствам керувати додатками в GCP, альтернативних хмарних постачальниках та локальних налаштуваннях [72].

Azure. Запущена у 2010 році, Microsoft Azure являє собою комплексну платформу хмарних обчислень, що пропонує різноманітний набір послуг для розробників та підприємств. Azure – одна з провідних хмарних платформ, дає компаніям гнучкість, масштабованість та надійність для розробки, реалізації та обслуговування сервісів і додатків. Привабливість платформи ще більше покращується завдяки інтеграції з широкою множиною корпоративних продуктів та сервісів від Microsoft, що дозволяє згладжувати взаємодію та дозволяє підприємствам максимально ефективно використовувати власні поточні інвестиції в технології Microsoft. Завдяки відданості до безпеки, узгодженості та можливостям гібридної хмари, Microsoft Azure є найкращим засобом для компаній в різних секторах, що бажають використовувати потенціал хмарних обчислень, водночас зберігаючи контроль над інфраструктурою та даними. Подальший розвиток Azure, пропонуючи абсолютно нові рішення для задоволення мінливих потреб сучасних компаній та розробників у цифрову добу, є доказом її адаптивності та креативності. Azure досягла значного прогресу в сфері ШІ та машинного навчання за рахунок Azure Machine Learning and Cognitive Services, що пропонує апріорі створені моделі ШІ та інструменти для створення власних моделей. Великі дані та сховища даних поєднуються для отримання інформативних результатів за рахунок впровадження Azure Synapse Analytics. Також, бажання Azure до квантових обчислень з Azure Quantum ставить її як лідера в дослідженні новітніх технологій обчислень [72].

Крім того пропонується розробка еталонної моделі для прийняття рішень, що ґрунтується на хмарних технологіях. Вона складається з п’яти складників [73].

1. **Хмарні системи, керовані даними**. За рахунок сховищ та послуг на вимогу з оплатою за використання, вони є непоганим варіантом для зберігання величезних обсягів даних. Завдяки можливості звітності щодо керування, аналізу даних та виявлення винятків, прикладне програмне забезпечення, що візуалізує ці дані, допомагає користувачам приймати рішення. Дослідження кореляції між відповідними результатами та незалежними факторами може надати допомогу у розробці гіпотез та ментальних моделей. Додатковими перевагами масової паралельної обробки у хмарних кластерах є ефективність та швидкість візуалізації.

2. **Підтримка рішень на основі моделей**. Використання математичних та інших типів моделей робить це можливим. Електронні таблиці дають можливість користувачам розробляти математичні моделі, виконувати аналіз “що, якщо”, і допомагають в прийнятті рішень. Наприклад, керівник страхової компанії може сформувати питання: “Як зміняться страхові виплати для страхового клієнта, якщо він змінить місце роботи?” Дослідницький внесок у сфері дослідження операцій та науки про управління є невід’ємними складовими для прийняття рішень на основі моделей. Окремим варіантом є розгортання автоматизованих алгоритмів прийняття рішень.

3. **Підтримка рішень на основі знань**. Можливість системи рекомендувати курс дій для менеджерів є однією з головних відмінностей цієї складової порівняно з іншими. Ці системи можуть вирішувати певні проблеми. Знання в певній сфері, усвідомлення проблеми у межах даної області та вміння вирішувати деякі з цих проблем задають функціональну множину.

4. **Підтримка рішень на основі документів**. Завдяки моніторингу, зберіганню, архівації та знаходження документів на основі ключових слів, індексації та змісту, дана система розширює можливості системи керування документами, що були створені з метою допомогти компаніям позбавитися паперового документообігу. Підприємства можуть отримати доступ до великого об’єму тексту зі згадками, коментарями, оглядів своїх продуктів та інших подій завдяки швидкому розвитку інтернету та соціальних мереж. Кілька

технологій, що можуть шукати в цьому величезному обсязі неструктурованих даних для пошуку релевантних документів, надають допомогу в прийняття рішень на основі документів. Особи, що приймають рішення, можуть слухати своїх клієнтів, використовуючи автоматизовані висновки, категоризацію, sentiment analysis, natural language processing, бібліометричний та лексичний аналіз. Впровадження оптичного розпізнавання символів є доволі простим.

5. **Підтримка рішень на основі комунікацій**. Даний елемент бере початок у дослідженні систем підтримки групових рішень. Складова включає в собі технології комунікацій, співпраці та підтримки рішень, які групи людей використовують разом. Хмарні системи добре підходять для підтримки рішень на основі комунікацій через необхідність координації, співпраці, планування, оцінки, прийняття рішень, досягнення згоди між членами команди в географічно розподілених глобальних організаціях. Такі системи дозволяють користувачам використовувати одночасно комунікації як моделі прийняття рішень.

Висновки до розділу 4

Розроблено систему підтримки прийняття рішень, що дозволяє обчислити майбутні страхові виплати на основі даних страхових клієнтів. Програма обчислює грошову суму за допомогою моделей УЛМ та ймовірності за допомогою мережі Байєса з використанням LS-методу. В якості даних використовувалися датасет клієнтів компанії Singapore Actuarial Society. Результати роботи ймовірнісної моделі було порівняно з розрахунками, отриманими за допомогою логістичної регресії. У більшості випадків мережа Байєса показала кращі результати. В основі УЛМ використовувався експоненційний розподіл з логарифмічною функцією зв’язку, нормальний розподіл з функцією зв’язку та розподіл Лапласа з тотожною функцією зв’язку. На основі критерії адекватності моделей було виявлено, що експоненційна УЛМ продемонструвала найкращі результати. Обрана регресійна модель і

мережа Байєса використовувалися для обчислення страхових виплат, що виступають в ролі ризику втрат для страхової компанії. Результати досліджень продемонстрували, що створена СППР доволі детально обчислює ризик втрат для страхової компанії. Розглянуто випадок масштабування СППР в різних контекстах. У подальших дослідженнях пропонується створити веб-версію додатку з застосуванням підходу REST API з можливістю додавання, оновлення або видалення даних страхових клієнтів. Разом з цим розглядається варіант застосування хмарних платформ.

ВИСНОВКИ

У дисертаційному дослідженні розв’язано задачу побудови інтелектуальної системи підтримки прийняття рішення для аналізу актуарних ризиків шляхом використання узагальнених лінійних моделей та підходу Байєсівського ймовірнісного оцінювання.

За результатами дисертаційної роботи можна зробити такі висновки.

1. Проведено аналіз існуючих методів оцінювання ризику, підходів стосовно аналізу актуарних ризиків. Було виявлено, що недостатньо уваги приділяється методам якісного оцінювання і виникає потреба у застосуванні економічних та математичних методів моделювання та аналізу фінансових ризиків та у створенні актуарних моделей з метою пошуку оптимального значення страхових премій за обчисленою ймовірністю банкрутства.

2. Побудовано моделі для оцінювання страхових ризиків у формі узагальнених лінійних моделей, Байєсівських мереж і логістичної регресії.

3. Вперше побудовано та досліджено вибрані типи узагальнених лінійних моделей для аналізу актуарних процесів, які відрізняються можливістю їх застосування для формального опису лінійних і нестаціонарних нелінійних процесів і забезпечують можливість оцінювання актуарного ризику з високою якістю.

4. Удосконалено метод оцінювання параметрів узагальнених лінійних моделей завдяки комплексному застосуванню методу Монте-Карло для марківських ланцюгів, методу адаптивного оцінювання моментів (Adam) та ітераційно рекурсивно зважувального методу найменших квадратів, що гарантує підвищення якості параметрів. У подальшому використовувався метод Монте-Карло для марківських ланцюгів, оскільки він здатний до адаптації у разі надходження нових даних.

5. Вперше побудовано нову модель на основі LS-методу у формі байєсівської мережі для оцінювання актуарних ризиків, яка характеризується коректністю формального опису вибраного процесу за байєсівським

інформаційним критерієм і забезпечує оцінювання високоякісних прогнозів актуарних процесів. Результати ймовірнісного висновку були порівняні з результатами логістичної регресії і було показано, що Байєсівська мережа у більшості випадків продемонструвала набагато кращі результати.

6. Створено систему підтримки прийняття рішень для оцінювання страхових ризиків.

7. Підготовлено статистичні набори даних, на основі яких було виконано ряд обчислювальних експериментів для оцінювання ризику у страховому сегменті.

8. Вперше створено і реалізовано програмно системну методологію побудови адаптивних моделей актуарних процесів, яка відрізняється запропонованим методом моделювання в умовах наявності статистичних невизначеностей даних, методом оцінювання параметрів моделей і забезпечує підвищення адекватності побудованих моделей та якості оцінок прогнозів.

9. Усі теоретичні розробки, отримані автором, доведені до конкретних моделей, інженерних методик, алгоритмів та засобів прогнозування ризиків у страхуванні, що дають можливість забезпечити функціонування СППР в умовах її постійної адаптації до ризикових ситуацій, зумовлених як зовнішніми так і внутрішніми факторами.

10. Результати дисертаційного дослідження впроваджено в Інституті телекомунікацій і глобального інформаційного простору Національної академії наук України з метою автоматизованого розв’язання задач моделювання, аналізу та прогнозування втрат страхових компаній.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Панібратов, Р. "Система підтримання прийняття рішень для оцінювання та прогнозування стану страхової компанії." *System research and information technologies,* no. 1, 2022, pp. 61-72. https://doi.org/10.20535/SRIT.2308-8893.2022.1.05.

2. Panibratov, R., and P. Bidyuk. "Estimation of the Parameters of Generalized Linear Models in the Analysis of Actuarial Risks." *System research and information technologies*, no. 2, 2023, pp. 139-148. https://doi.org/10.20535/SRIT.2308-8893.2023.2.10.

3. Panibratov, R. S. "Analysis of data uncertainties in modeling and forecasting of actuarial processes." *Radio Electronics, Computer Science, Control*, no. 2, 2024, pp. 45-51. https://doi.org/10.15588/1607-3274-2024-2-5.

4. Panibratov, R. S, and P. I. Bidyuk. “Analysis of actuarial risk with generalized linear models.” *System Research and Information Technologies*, no. 4, 2025, pp. 58-70. https://doi.org/10.20535/SRIT.2308-8893.2025.4.04.

5. Panibratov, R., and P. Bidyuk. "Applying Bayesian networks in analysis of actuarial risks." *International Scientific Technical Journal" Problems of Control and Informatics"*, vol. 70, no. 3, 2025, pp. 33-44. https://doi.org/10.34229/1028-0979-2025-3-3.

6. Panibratov, R. “Methods of estimating parameters of generalized linear models in the analysis of actuarial risks.” *Proceedings of V international scientific and methodical conference Computer technology and mechatronics*, Kharkiv, Ukraine, 24 May 2023, pp. 86-89. https://dl2022.khadi-kh.com/mod/resource/view.php?id=361231.

7. Panibratov, R. "Analysis of data uncertainties in modeling and forecasting of actuarial processes." *Proceedings of 3rd International Scientific and Practical Internet Conference Scientific Research and Innovation*, Dnipro, Ukraine, 18-19 April 2024, edited by V. V. Marenichenko, WayScience, pp. 19-21.

http://www.wayscience.com/wp-content/uploads/2024/04/Conference-Proceedings-April-18-19-2024-1.pdf.

8. Panibratov, R. “Analysis of actuarial risk with generalized linear models.” *Proceedings of 6th International Scientific and Practical Internet Conference Integration of Education, Science and Business in Modern Environment: Winter Debates*, Dnipro, Ukraine, 6-7 February 2025, edited by V. V. Marenichenko , WayScience, pp. 37-39. http://www.wayscience.com/wp-content/uploads/2025/02/Conference-Proceedings-February-6-7-2025-1.pdf.

9. Panibratov, R., V. Gavrylenko, and P. Bidyuk. “The system analysis based approach to market and credit risks analysis.” *Матеріали III Всеукраїнської наукової конференції здобувачів освіти і молодих учених Відбудова транспортної інфраструктури України*, Київ, Україна, 25 березня 2025 року, pp. 252-253. https://doi.org/10.33744/978-966-632-331-9-2025-1.

10. Panibratov, R. S. “Applying Bayesian networks in analysis of actuarial risks.” *Proceedings of 1st International Scientific and Practical Internet Conference “Crossroads of Ideas: Science, Technology and Society in a Global Context”*, Dnipro, Ukraine, 10-11 April 2025, edited by V. V. Marenichenko , WayScience, pp. 7-9. http://www.wayscience.com/wp-content/uploads/2025/04/Conference-Proceedings-April-10-11-2025.pdf.

11. Слободянюк, О. В., and І. Ф. Примаченко. "Страховий ринок: драйвер економічної безпеки." *Науковий вісник Львівського державного університету внутрішніх справ (серія економічна)*, no. 1, 2023, pp. 88-94. https://doi.org/10.32782/2311-844X/2023-1-12.

12. Житар, М. "Тенденції розвитку страхового ринку України в умовах воєнного стану." *Економіка та Суспільство*, no. 61, 2024, pp. 1-6. https://doi.org/10.32782/2524-0072/2024-61-24.

13. Майстер, А. В., and А. А. Славкова. "Страховий ринок України: проблеми та перспективи розвитку." *Інвестиції: практика та досвід*, no. 2, 2025, pp. 108-113. https://doi.org/10.32702/2306-6814.2025.2.108.

14. Братюк, В. П. "Сучасний стан страхового ринку в Україні." *Таврійський науковий вісник. серія: економіка*, no. 12, 2022, pp. 37-45. https://doi.org/10.32851/2708-0366/2022.12.5.

15. Череп, А. В., and Ю. П. Кішко. "Перспективи розвитку страхового ринку України в умовах євроінтеграційних процесів." *Економіка та підприємництво: Держава та регіони*, no. 2 (132), 2024, pp. 78-83. https://doi.org/10.32782/1814-1161/2024-2-13.

16. Черкасова, С. В., and А. С. Бойчук. "Сучасні тренди розвитку страхового ринку та чинники забезпечення його фінансової безпеки." *Вісник ЛТЕУ. Економічні науки*, no. 81, 2025, pp. 59-65. https://doi.org/10.32782/2522-1205-2025-81-08.

17. Захарченко, Н. В., and С. А. Явір. "Вплив COVID-19 на ринок страхування в Україні та у всьому світі." *Ринкова економіка: сучасна теорія і практика управління*, vol. 21, no. 4 (47), 2021, pp. 110-128. https://doi.org/10.18524/2413-9998.2021.1(47).227009.

18. Горбачова, О., and А. Петрова. "Цифровізація страхового ринку України: основні проблеми та перспективи в умовах воєнного стану." *Сталий розвиток економіки*, no. 2 (53), 2025, pp. 29-36. https://doi.org/10.32782/2308-1988/2025-53-4.

19. Вовк, В., Ю. Жежерун, and В. Костогриз. "Страховий ринок України у період дії воєнного стану: фінансовий та маркетинговий аспекти." *Проблеми і перспективи економіки та управління*, no. 3 (35), 2023, pp. 119-131. https://doi.org/10.25140/2411-5215-2023-3(35)-119-131.

20. Добрик, Л., М. Абрамов, and І. Рябінін. "Воєнний вплив на страховий ринок України: аналіз перешкод та реструктуризація фінансового ландшафту." *Review of transport economics and management*, no. 10 (26), 2023, pp. 218-229. https://doi.org/10.15802/rtem2023/292851.

21. "Обсяги активів зросли у фінансових компаній, страховиків і ломбардів — Огляд небанківського фінансового сектору." *Національний банк України,* 31 березня 2025. https://bank.gov.ua/ua/news/all/obsyagi-aktiviv-zrosli-u-

finansovih-kompaniy-strahovikiv-i-lombardiv--oglyad-nebankivskogo-finansovogo-sektoru.

22. Aquilina M., et al. “The transformation of the life insurance industry: systemic risks and policy challenges.” *BIS Papers*, 2025.

23. Трухан, С. В., and П. І. Бідюк. "Застосування мереж Байєса до побудови моделей оцінювання ризику актуарних процесів." *ScienceRise*, vol. 8, no. 2 (25), 2016, pp. 6-14. https://doi.org/10.15587/2313-8416.2016.74962.

24. Tripp, M. H, et al. "Quantifying operational risk in general insurance companies. Developed by a Giro working party." *British Actuarial Journal*, vol.10, no. 5, 2004, pp. 919-1012. https://doi.org/10.1017/S1357321700002919.

25. Shah, S. “Measuring Operational Risk Using Fuzzy Logic Modeling”, *IRMI*, 1 Sep. 2003, https://www.irmi.com/articles/expert-commentary/measuring-operational-risk-using-fuzzy-logic-modeling.

26. “International Convergence of Capital Measurement and Capital Standards. A Revised Framework. Comprehensive Version.” *Bank for Int. Settlements*, 1 Jun. 2006, https://www.spglobal.com/content/dam/spglobal/mi/en/documents/general/Risk-Insight-Internal-Default-Risk-Models-Alive-And-Kicking.pdf.

27. “Sound Practices for the Management and Supervision of Operational Risk.” *Bank for Int. Settlements*, 21 Dec. 2010. https://www.bis.org/publ/bcbs183.pdf.

28. Medova, E. "Extreme value theory: extreme values and the measurement of operational risk." *Operational Risk*, vol. 1. no. 7, 2000, pp. 13-17.

29. *“*Scenario-based AMA”*, Federal reserve Bank of New York,* 1 May 2003, https://www.newyorkfed.org/medialibrary/media/newsevents/events/banking/2003/con0529d.pdf .

30. Vanini, P., M. Leippold, and B. Döbeli. “From Operational Risk to Operational Excellence.” *SSRN Electronic Journal*., 2003, pp. 1-18 https://doi.org/10.2139/ssrn.413720.

31. Kühn, R., and P. Neu. "Functional correlation approach to operational risk in banking organizations." *Physica A: Statistical Mechanics and its Applications*, vol 322, 2003, pp. 650-666. https://doi.org/10.1016/S0378-4371(02)01822-8.

32. Nguyen, H. T., C. Walker, and E. A. Walker. “A first course in fuzzy logic.”, *Chapman and Hall/CRC*, 2018.

33. Popescu, V. F., and M. S. Pistol. "Fuzzy logic expert system for evaluating the activity of university teachers." *International Journal of Assessment Tools in Education*, vol. 8, no. 4, 2021, pp. 991-1008. https://doi.org/10.21449/ijate.1025690.

34. Бідюк П. І., et al. “Системи і методи підтримки прийняття рішень: навчальний посібник.” *НТУУ "КПІ"*, 2022.

35. “Top Actuarial Technologies of 2022-2023”, *Society of actuaries*, 1 Sep. 2023, https://www.soa.org/4a7ea2/globalassets/assets/files/resources/research-report/2023/top-actuarial-tech-2022-2023.pdf.

36. “An Actuarial take on Enterprise Risk Management”, *PricewaterhouseCoopers Luxembourg*, 1 Sep. 2023, https://blog.pwc.lu/an-actuarial-take-on-enterprise-risk-management/.

37. “Actuarial Modelling Centre”, *Deloitte*, 1 Feb. 2026, https://www.deloitte.com/ro/en/services/consulting/services/actuarial-modelling-centre.html?icid=mosaic-grid_actuarial-modelling-centre..

38. Згуровський, М. З. et al. Байєсівські мережі в ситемах підтримки прийняття рішень(навчальний посібник).” *Едельвейс*, 2015.

39. Lauritzen S. L., and D. J. Spiegelhalter. "Local computations with probabilities on graphical structures and their application to expert systems." *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 50, no.2, 1988, pp. 157-194. https://doi.org/10.1111/j.2517-6161.1988.tb01721.x.

40. Wang, B., Y. Chen, and Z. Li. "A novel Bayesian Pay-As-You-Drive insurance model with risk prediction and causal mapping." *Decision Analytics Journal*, vol. 13, 2024, pp.1-8. https://doi.org/10.1016/j.dajour.2024.100522.

41. Haines N., C. Goold, and J. M. Shoun. "A Bayesian workflow for securitizing casualty insurance risk", *arXiv*, 9 Jul. 2025. https://arxiv.org/pdf/2407.14666.

42. Ayyadurai R., et al. "Exploring the synergy of self-supervised learning and Bayesian networks for customer profiling in the insurance industry." *International Journal of Automation and Smart Technology*, vol. 15, no. 1, 2025, pp. 1-14. DOI: https://doi.org/10.5875/py039m26.

43. D. Anderson et al. "A Practitioner's Guide to Generalized Linear Models – a foundation for theory, interpretation and application; 3rd edition." *Towers Watson*, 2007.

44. McCullagh P. and J. Nelder. "Generalized Linear Models; 2nd edition." *Chapman & Hall*, 1989.

45. Марчук В. В. "Використання моделей з нечіткою логікою для прийняття інвестиційних рішень щодо матеріальних активів підприємства." *Збірник наукових праць "Вчені записки"*, no. 39 (2), 2025, pp. 8-23. DOI: http://doi.org/10.33111/vz_kneu.39.25.02.01.005.011.

46. Al-Nahhas, Y. S., et al. "Modified Mamdani-fuzzy inference system for predicting the cost overrun of construction projects." *Applied Soft Computing*, vol. 151, 2024, pp. 1-23. https://doi.org/10.1016/j.asoc.2023.111152.

47. Шовгун, Н. В. "Аналіз кредитоспроможності позичальника за допомогою методів з нечіткою логікою." *Вісник Національного технічного університету України КПІ. Іформатика, управління та обчислювальна техніка*, no. 54, 2011, pp. 224-228.

48. Toma, S. V., M. Chiriță, and D. Șarpe. "Risk and uncertainty." *Procedia Economics and Finance*, vol. 3, 2012, pp. 975-980. https://doi.org/10.1016/S2212-5671(12)00260-2.

49. Brijith, A. "Data preprocessing for machine learning." *Int. Center AI Cyber Security Res. Innovations*, vol. 3, 2023, pp.1-4.

50. Bolikulov, F., et al. "Effective methods of categorical data encoding for artificial intelligence algorithms." *Mathematics*, vol., 12, no.16, 2024, pp. 1-21.

51. Chen, X., and J. C. Jeong. "Enhanced recursive feature elimination." *Proceedings of sixth international conference on machine learning and applications (ICMLA 2007)*. IEEE, 2007, pp. 429-435. https://doi.org/10.1109/ICMLA.2007.35.

52. McHugh, M. L. "The chi-square test of independence." *Biochemia medica*, vol. 23, no. 2, 2013, pp. 143-149. https://doi.org/10.11613/BM.2013.018.

53. Alakkari, K., and B. Ali. "Using Random Forest Algorithm to Improve Investment Decision Making in Damascus Stock Exchange." *EDRAAK*, vol. 2025, 2025, pp. 57-61. https://doi.org/10.70470/EDRAAK/2025/008.

54. Urrea, C., and R. Agramonte. "Kalman filter: historical overview and review of its use in robotics 60 years after its creation." *Journal of Sensors*, vol. 2021, no. 1, 2021, pp. 629-638. https://doi.org/10.1155/2021/9674015.

55. Клюєв О. et al. "Використання фільтра Калмана як спостерігача стану у векторній системі керування асинхронною машиною.", *Вісник КрНУ імені Михайла Остроградського*, vol. 5 (136), 2022, pp. 27-35. DOI: https://doi.org/10.32782/1995-0519.2022.5.3.

56. Garćia, L. P. F., A. C. de Carvalho, and A. C. Lorena. "Noisy data set identification." Proceedings of *International Conference on Hybrid Artificial Intelligence Systems*. Springer Berlin Heidelberg, 2013, pp. 629-638. https://doi.org/10.1007/978-3-642-40846-5_63.

57. Zhu, X., and X. Wu. "Class noise vs. attribute noise: A quantitative study." *Artificial intelligence review*, vol. 22, no. 3, 2004, pp. 177-210. https://doi.org/10.1007/s10462-004-0751-8.

58. Wu, X. "Knowledge acquisition from databases.", *Ablex Publishing Corp*, 1996.

59. Dibal, N. P, R. Okafor, and H. Dallah. "Challenges and implications of missing data on the validity of inferences and options for choosing the right strategy in handling them." *International Journal of Statistical Distributions and Applications,* vol 3, no. 4, 2017, pp. 87-94. https://doi.org/10.11648/j.ijsd.20170304.15.

60. Graham, J. W., P. E. Cumsille, and E. Elek-Fisk. "Methods for handling missing data." *Handbook of psychology*, 2003, pp. 87-114. https://doi.org/10.1002/0471264385.wei0204.

61. He, Y. "Missing data analysis using multiple imputation: getting to the heart of the matter." *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, no. 1, 2010, pp. 98-105. https://doi.org/10.1161/CIRCOUTCOMES.109.875658.

62. Бідюк, П.І., О. Гожий, and М.М. Коновалюк. "Прогнозування волатильності валютного ринку за нелінійними моделями." *Вісник Національного університету Львівська політехніка*, no. 719, pp. 154 – 163.

63. Affet-Sahalia, Y. "Transition Densities for Interest Rate and Other Nonlinear Diffusions." *The Journal of Finance*, vol. 54, no. 4, 1999, pp. 1361–1395. https://doi.org/10.1111/0022-1082.00149.

64. Zhang, T. "Iteratively reweighted least squares with random effects for maximum likelihood in generalized linear mixed effects models." *Journal of Statistical Computation and Simulation*, vol. 91, no. 16, 2021, pp. 3404-3425. https://doi.org/10.1080/00949655.2021.1928127.

65. Kingma, D. P, and J. Ba. "Adam: A method for stochastic optimization.", arXiv, 30 Jan. 2017, https://arxiv.org/pdf/1412.6980.

66. Kavianihamedani, H., J. D. Quinn, and J. D. Smith. "New diagnostic assessment of MCMC algorithm effectiveness, efficiency, reliability, and controllability." *IEEE Access,* vol. 12, 2024, pp. 42385-42400. https://doi.org/10.1109/ACCESS.2024.3378752.

67. Hörmann, W., and J. Leydold. Monte Carlo integration using importance sampling and Gibbs sampling. *Proceedings of the International Conference on Computational Science and Engineering*. Istanbul, 1 Jan. 2005, pp. 92-97.

68. Hesterberg, T. "Weighted average importance sampling and defensive mixture distributions.", *Technometrics*, vol. 37, no. 2, 1995, pp. 185-194.

69. Al Luhayb, A. S. M. "The bootstrap method for Monte Carlo integration inference." *Journal of King Saud University-Science*, vol. 35, no.6, 2023, pp. 1-5. https://doi.org/10.1016/j.jksus.2023.102768.

70. Darwiche, A. “Modeling and reasoning with Bayesian networks.” *Cambridge University Press*, 2009. https://doi.org/10.1017/CBO9780511811357.

71. Abid, L. “A logistic regression model for credit risk of companies in the service sector.”, *International Research in Economics and Finance* , vol. 6, no. 2, 2022, pp. 1-10. https://doi.org/10.20849/iref.v6i2.1179.

72. Alkhatib, A, A. Shaheen, and R. N. Albustanji. "A comparative analysis of cloud computing services: AWS, Azure, and GCP.", *genesis*, vol. 4, 2024, pp.1-15. http://dx.doi.org/10.12785/ijcds/1571111846.

73. Kaparthi, S, M. Arti, and D. J. Power. "An overview of cloud-based decision support applications and a reference model." *Studies in Informatics and Control*, vol. 30, no. 1, 2021, pp. 5-18. https://doi.org/10.24846/v30i1y202101.

ДОДАТОК А ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ

ДОВІДКА
про впровадження результатів дисертаційного дослідження
Панібратова Романа Сергійовича
на тему «Система підтримки прийняття рішень для аналізу актуарних ризиків»

Результати, отримані в ході виконання дисертаційного дослідження Панібратова Романа Сергійовича на тему «Система підтримки прийняття рішень для аналізу актуарних ризиків» мають практичне значення, їх впровадження є доцільним. Зокрема, вони можуть бути використані у науковій та науково-дослідній роботі Інституту телекомунікацій і глобального інформаційного простору Національної академії наук України для розв'язання задач дослідження ризиків, розроблення методик їх оцінювання та визначення потенційних наслідків.

Зокрема, застосовуються результати дисертаційної роботи, пов'язані із побудовою моделей оцінювання та аналізу ризиків на основі статистичних і машинних методів, формуванням рекомендацій та підтримкою прийняття рішень в умовах невизначеності, використанням байєсівських підходів і методів імовірнісного моделювання, розробленням інформаційно-аналітичних компонентів для оброблення даних та прогнозування показників ризику.

Ця довідка не є підставою для фінансових розрахунків.

Директор Інституту
член-кореспондент НАН України
д.т.н., професор Олександр ТРОФИМЧУК