JIRSS (2025)
Vol. 24, No. 01, pp 35-66
DOI: 10.22034/jirss.2025.2033194.1067

A Model for Determining Insured Premiums Based on Household Expenses Using Advanced Computational Techniques under Heterogeneous Data Conditions

Amir Akbarzadeh Janatabad[1], **Ahmad Sadegheih**[1], **M. M. Lotfi**[1]

[1] Department of Industrial Engineering, Yazd University, Iran.

Received: 20/06/2024, Accepted: 31/08/2025, Published online: 16/11/2025

Abstract. The escalating public health costs are a significant concern for governments globally. The efficient management of those costs is critical, with health insurance systems playing a pivotal role. However, the insurance industry faces challenges due to the heterogeneous data, leading to inconsistent outputs for identical inputs. Traditional predictive methods such as Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) often fail to address these inconsistencies. This study proposes a novel two-stage model to determine insurance premiums, incorporating equity considerations and advanced computational techniques. We advocate for an expenditure-based premium calculation as a superior alternative to the traditional salary-based approach. This method aligns premiums more closely with household expenses, promoting fairness and efficiency. Our results demonstrate that the expenditure-based strategy outperforms the salary-based one in controlling costs for both the insured and the insurer. Specifically, the error metrics, including Mean Absolute Error and Root Mean Square Error, show significant improvement in our model compared to the ANFIS method. To enhance the model's accuracy, we integrate sampling techniques to mitigate the data heterogeneity and employ genetic algorithms to optimize the weights of the neural network. The genetic algorithm iteratively evolves the network parameters, ensuring robust performance even in diverse data. Our findings indicate that this integrated approach significantly reduces prediction errors and enhances the overall reliability of the premium calculation process. In conclusion, the proposed model offers a robust framework for premium determination, addressing the inherent data heterogeneity in the insurance industry. This study provides a valuable contribution to

Amir Akbarzadeh Janatabad (akbarzadeh.am@yahoo.com)
Corresponding Author: Ahmad Sadegheih (sadegheih@yazd.ac.ir)
M. M. Lotfi (lotfi@yazd.ac.ir).

the field by demonstrating a practical and effective solution for improving the accuracy and fairness of insurance premium calculations.

Keywords. Data Mining, Machine Learning, Fuzzy Logic, Health Insurance, Neural Network.

MSC: 62P25, 68T07, 62C86.

1 Introduction

According to Folland et al. (2016), health insurance supported by governments in many countries aims to reduce health-related costs while maintaining the quality of care. The health insurance system is intricately connected with the social insurance system and the broader health system, forming a core component of the social welfare structure. Its effectiveness largely depends on the comprehensiveness of the health insurance framework. Maintaining and improving citizen's health is considered a fundamental responsibility of any government and requires significant public funding. This has become a major concern for healthcare managers and decision-makers, who continuously seek strategies to optimize outcomes and control expenditures.

Voto and Ngepah (2025) note that addressing the financial burden of healthcare is essential for ensuring universal health coverage (UHC) and reducing financial barriers, particularly for low-income populations. Renowned as the most expensive country in the world for healthcare, the true cost of health insurance in the United States is a big talking point. Like other countries, the cost of health insurance in the US is based on several factors relevant to each individual.

In today's world, insurance serves as a crucial economic tool that significantly impacts the socio-economic conditions of individuals. Therefore, the pricing process in the insurance industry requires careful attention. Experts often argue that current pricing strategies are inadequate due to the lack of competitive conditions, calling for a robust framework for pricing various insurance contracts. Although the payment functions and conditions of insurance contracts resemble those of financial contracts, traditional financial pricing methods do not adequately address insurance risks because of their diverse distribution profiles. Effective insurance risk management is essential and must be prioritized within insurance companies.

Insurance pricing is inherently risky and requires cautious and informed decision-making. Implementing various pricing strategies enhances an organization's ability to effectively assess risks, thereby improving profitability and gaining insights into consumer behaviour. According to Timothy et al. (2017), poor choices in the health insurance market lead to inefficiencies such as adverse selection and inadequate service provision, resulting in consumer dissatisfaction. A critical aspect of controlling treatment costs within health insurance systems is the accurate pricing of premiums. Ideally, premiums should reflect the actual costs and damages incurred by the insured.

With healthcare expenditures becoming more complex, it is critical for insurance companies and policyholders to accurately estimate insurance prices. Utilising a dataset that includes medical history, demographic data, and other pertinent variables, a variety of machine learning techniques, such as ensemble methods and regression, are used to create prediction models. Patil et al. (2024).

Rose (2013) notes that for government employees, premiums are typically calculated annually based on salary, insurance type, and coverage level, without considering individual cost and claims records, which can be seen as a form of social injustice.

A significant challenge in achieving accurate premium pricing is data heterogeneity. In the insurance industry, data heterogeneity refers to the diverse nature of data inputs and outputs. For instance, similar inputs (such as demographic information and medical history) can lead to different claims and cost patterns, complicating prediction and pricing processes. This heterogeneity arises from various sources, including differing health conditions, treatment protocols, and regional cost variations, creating a complex landscape that traditional predictive models often struggle to navigate.

To address these challenges, this study proposes a novel two-stage model that integrates data mining (DM) techniques and fuzzy logic for enhancing premium determination. The first stage involves using DM techniques, such as clustering and regression analysis, to predict household health expenditures. This stage includes preprocessing activities to handle missing values, normalize data, and reduce dimensionality through methods like principal component analysis. In the second stage, fuzzy logic is employed to calculate premiums based on the predicted expenditures, ensuring a fairer premium distribution that rewards lower costs incurred by the insurer.

By focusing on an expenditure-based premium model, this research aims to provide a more equitable and efficient alternative to the traditional salary-based approach. This model addresses data heterogeneity by integrating advanced computational techniques, ensuring more accurate and reliable premium calculations, ultimately leading to better cost control for the insured and the insurer.

In today's world, data equates to information a key asset for organizations. Information can improve humanity's future, and this is the primary objective of scientists and researchers. However, heterogeneous data complicates the acquisition of meaningful insights. The data discussed in this article also complicates calculations. We encounter a volume of data defined as input variables in the model, which may be similar but produce different outputs. Consequently, when this data is processed using software, the results can be ambiguous and lack logical applicability. A viable solution needs to be found to address this challenge. Based on this discussion, the current work partly fills the gap in this field by proposing a suitable model for determining reasonable insurance premiums. If the premium charged to each household is based on an indicator such as household expenses, some responsibility for resource and cost management will shift to the insured (currently, the insurance premium is 7% of salary).

Therefore, the study poses the following questions:

1) How can household medical expenses be factored into premium determination?

2) Considering the heterogeneous data in the insurance industry, how should the required model be designed? (It is important to note that heterogeneous data creates challenges in decision-making.)

If insurance premiums are based on household expenses, it benefits both insurance companies and the insured. Individuals who impose lower costs on insurance companies would pay lower premiums.

2 Literature Review

Many articles in this field were reviewed to achieve the research goal of pricing insurance services within the insurance industry.

2.1 Traditional Insurance Pricing Models

Traditionally, insurance companies have relied on deterministic models and actuarial science principles to determine premiums. Pantelous and Passalidou (2013) introduced a dynamic programming-based approach to optimize premium pricing policies in competitive environments. While their model is sophisticated, it does not incorporate emerging technologies such as artificial intelligence and soft computing. This area of research emphasizes the historical dependence on deterministic methodologies and the necessity for more advanced techniques to tackle contemporary challenges in the insurance industry.

2.2 Integration of AI and Soft Computing

In response to the limitations of traditional pricing models, researchers have increasingly focused on artificial intelligence (AI) and soft computing techniques. Studies like those by Goodarzi and Janat Babaei (2016) and Boodhun and Jayabalan (2018) have investigated the application of machine learning (ML) algorithms and artificial neural networks (ANN) in the insurance sector. These advanced computational methods provide greater flexibility and adaptability, enabling insurers to better address the complexities of risk assessment and premium determination in dynamic environments. This body of literature highlights the growing recognition of AI and soft computing as essential tools for modernizing insurance pricing practices.

2.3 Importance of DM in Pricing Strategies

The importance of data mining (DM) techniques in insurance pricing strategies cannot be overstated. DM allows insurers to extract valuable insights from large and

diverse datasets, leading to more informed decision-making and better risk management. Studies by Brofer et al. (2017) and Rezaei Navaei and Koosha (2016) have shown the effectiveness of DM in identifying patterns, predicting claim behaviour, and optimizing premium structures. By utilizing DM techniques, insurers can strengthen their competitive advantage and enhance profitability. This section highlights the crucial role of data-driven approaches in shaping modern pricing strategies within the insurance industry.

2.4 DM Techniques

Wanke and Barros (2016) examined the role of heterogeneity in the insurance sector and emphasized that data mining (DM) methods may lead to two types of errors: incorrectly assigning high-efficacy observations to a low-efficiency group and vice versa. To clarify these terms:

High-efficacy Observation: An observation that demonstrates a high degree of accuracy or effectiveness in predicting or influencing the output variable. In the context of insurance, this might refer to data points that accurately depict high-risk or high-cost situations, where the predictions or classifications align closely with the actual outcomes.

Low-efficacy Observation: An observation with a lower degree of accuracy or effectiveness regarding predicting or influencing the output variable. This may encompass data points where the predictions or classifications fail to accurately capture the true risk or cost associated with the observation.

High-efficiency Group: A classification group where the observations exhibit high predictive performance and low error rates when utilizing DM methods. This group is characterized by its ability to accurately predict or classify the target variable, resulting in more reliable and valid outcomes.

Low-efficiency Group: A classification group characterized by lower predictive performance and higher error rates. Observations in this group often lead to less accurate predictions and may result in greater discrepancies between the predicted and actual outcomes.

The issues raised by Wanke and Barros highlight the importance of properly managing and selecting DM techniques to mitigate such errors. Techniques such as Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANN), Classification and Regression Trees (CART), k-nearest Neighbors (KNN), Bagging, and others were analyzed through comparisons of Mean Squared Error (MSE) metrics, with graphical representations of their results. The effectiveness of these techniques can vary based on the data's nature and the specific application, necessitating careful consideration and control over the chosen methods.

Furthermore, some researchers have explored the integration of DM techniques with advanced methods like artificial bee colonies and fuzzy logic to tackle challenges in

the insurance industry. These integrated solutions aim to enhance prediction accuracy and efficiency by leveraging the strengths of various computational approaches.

2.5 Combining DM Techniques with other Algorithms

Yan et al. (2019) highlighted the growing issue of car insurance fraud on a global scale. They introduced the Kernel Ridge Regression (KRR) optimized by the artificial bee algorithm (ABC), resulting in the KRR-ABC algorithm for fraud detection in car insurance. The performance of the KRR-ABC model was benchmarked across eight datasets and compared with other methodologies. The results indicated that the model outperformed others in performance and execution speed.

To address their issue, Wang and Liaw (2020) introduced evolutionary multitasking optimization (EMO) for fuzzy genetic data extraction, utilizing a genetic algorithm (GA) as an advanced genetic data mining method. They based their approach on a structure representation of the well-known multifactorial evolutionary algorithm (MFEA). Various tests demonstrated that their method enhanced the structure-based GA in terms of solution quality and convergence speed across all dataset sizes. In terms of evaluation efficiency, their proposed method was approximately 21 times faster than the structure-based GA.

Kalra et al. (2022) automated the assessment of insurance claims using various data techniques, which resulted in increased company credibility and customer satisfaction. Panda et al. (2022) developed a health insurance price prediction system called MLHIPS using machine learning algorithms, facilitating insurance companies in determining premium values and health costs quickly and efficiently through a multinomial regression model.

Özgur and Yolcu (2023) focused on estimating premium production for several insurance companies in Turkey using artificial neural networks (ANN), analyzing different training algorithms and transfer functions. Jones and Swati (2023) explored the application of machine learning techniques and big data analytics within the insurance sector, demonstrating the development of various predictive models such as AdaBoost, Naive Bayes, K-Nearest Neighbors (KNN), and decision trees.

Goel and Chaudhary (2024) employed advanced machine learning techniques, including linear regression and ANN, to uncover key insights into the impact of smoking habits and age on healthcare costs, equipping the industry with practical tools to enhance the accuracy of healthcare pricing.

Wilson et al. (2024) investigated insurance pricing in the motor insurance sector, evaluating multiple machine learning methods, including generalized linear models (GLM), gradient boosting machines (GBM), ANN, and a unique hybrid model that combines GLM and ANN. Upon review, the hybrid model demonstrated superior performance compared to the others, with ANN predictions closely matching the

performance of the combined model.

Explainable Artificial Intelligence (XAI) in insurance focuses on making the decision-making processes of AI models transparent and understandable, particularly for insurance related tasks like underwriting, claims processing, and fraud detection. XAI helps insurers build trust, ensure accountability, and comply with regulatory requirements by providing insights into how AI models arrive at their conclusions. **?**.

Abdi et al. (2017) note that a multi-stage data mining approach can effectively address customer insurance coverage sales plan problems by identifying high-potential loyal customers and predicting their likelihood of purchasing specific coverage. This approach involves several stages, including data cleansing, feature selection, clustering, and classification, to identify customer segments and predict insurance needs. The first stage addresses data cleansing. In the second stage, several filter and wrapper methods are implemented to select proper features. In the third stage, K-nearest neighbor algorithm is used to cluster the customers. The approach aims to select a compact feature subset with the maximal prediction capability. The proposed approach can detect the customers who are more likely to buy a specific insurance coverage at the end of a contract term.

It is important to note that heterogeneous data presents challenges and increases the likelihood of errors. Although the studies above have yielded positive results in the insurance industry, acceptable outcomes cannot be achieved with heterogeneous data, such as those used in the current study. To address this issue, we reviewed additional research in the field. Kumar Dubey et al. (2012) proposed an effective technique for knowledge discovery using a superset and subset approach. A frequent superset encompasses more transactions than the minimum support, while a frequent subset involves fewer transactions than the minimum support. This method enables the identification of increased connectivity, representing the most acceptable connections among sets of items.

The methodology employed in this study was dynamic. Kalra et al. (2018) utilized the K-Mean clustering algorithm on real heterogeneous datasets and analyzed the results based on cluster types. In most studies, data that is not uniform is referred to as heterogeneous. For example, Salama et al. (2022) focused on developing an ensemble machine learning model with active learning to identify the most effective feature extraction methods for detection and to perform heterogeneous data analysis in comparison with traditional machine learning algorithms. The proposed model was evaluated throughout the experimentation process and involved five heterogeneous datasets from various domains, including the Healthcare Reform dataset, Financial Phrase Bank dataset, Sander Frandsen dataset, SMS spam dataset, and Textbook Sales dataset. They analyzed the data of internet users and incorporated human experts alongside machine learning techniques.

Ghuse et al. (2017) performed logistic regression classification and implemented random forest and an improved artificial neural network to enhance the detection of

health insurance fraud, emphasizing the importance of data partitioning in classification algorithms. Other studies did not focus on reducing data heterogeneity. Since such data is prevalent in the insurance industry, this study is particularly relevant. The paper examines variables affecting the output at each stage of the analysis; for instance, it investigates how salary and the number of purchased insurance policies influence household spending. Notably, no previous research has explored the impact of these variables. Relying solely on forecasting methods such as neural networks or adaptive neuro-fuzzy inference systems does not satisfy our requirements. We aim to estimate premiums to ensure fairness.

The methods used in this study such as sampling, variable reduction, and the role of GA significantly contribute to the heterogeneity of the data.

2.6 Research Gap and Contribution

The literature review clarifies various approaches and methodologies used in insurance premium pricing. A critical analysis, however, reveals a significant research gap regarding the integration of advanced data mining (DM) techniques and the reduction of data heterogeneity in the context of determining insurance premiums. While some existing studies have examined the application of DM methods, few have thoroughly addressed the challenges posed by heterogeneous data and its effect on the accuracy of premium estimation. This study seeks to bridge this gap by proposing a novel methodology that combines advanced DM techniques with strategies to lessen data heterogeneity.

The learning process for all machine learning (ML) techniques can be disrupted when confronted with a dataset that produces different outputs for the same input data. In instances where it is not feasible to remove all records with the same input variables, our research method becomes beneficial, representing a significant innovation. Other studies that have employed DM techniques or addressed heterogeneous data have not tackled this specific challenge, which is inherently present in the insurance industry.

Our contribution is the development of a comprehensive framework that not only utilizes ML algorithms for predictive modelling but also integrates sampling techniques and genetic algorithms (GA) to tackle data heterogeneity and enhance model robustness. For instance, by categorizing the data into different classes, sampling from each class, and applying appropriate learning techniques specific to those classes, we significantly contribute to the literature on this subject while addressing the issue of data heterogeneity.

Contribution is paramount, as accurate premium estimation is critical for insurance companies to manage risk effectively and ensure financial stability. By incorporating DM techniques and addressing data heterogeneity, our methodology provides insurers with a more accurate and reliable means for pricing premiums, thereby improving decision-making processes and enhancing the overall efficiency and profitability of the

Table 1: Summary of Relevant Literature.

Researchers	Year	Selling products & customer focus	Damage to insurer	Fraud in insurance	Other industries (e.g., disease diagnosis)	DM techniques	Combined DM	DM + other techniques (e.g., PSO)	Data heterogeneity
Goodarzi & Janat Babaei	2016			✓		✓			
Wanke & Barros	2016	✓				✓			
Broffer et al.	2017	✓						✓	
Abdi et al.	2017	✓				✓			
Cigsar & Unul	2019		✓			✓			
Amini & Abdi	2018				✓	✓			
Boodhun & Jayabalan	2018	✓				✓			
Rezaei & Koosha	2016	✓						✓	
Yan et al.	2019			✓				✓	
Reddy et al.	2019				✓			✓	
Debashish et al.	2016			✓		✓			
Yin et al.	2021		✓			✓			
Zhao et al.	2021			✓				✓	
Sharma	2021			✓		✓			
Wang & Liaw	2020			✓				✓	
Lanjewar & Chaudhary	2020			✓		✓	✓		
Raja & Pandian	2020			✓				✓	
Hanafy & Ming	2021	✓				✓			
Panda et al.	2022	✓				✓			
Goel & Chaudhary	2024	✓				✓			
Wilson et al.	2024	✓						✓	
Jones & Swati	2023	✓		✓				✓	
Current study	**2024**	✓	✓			✓	✓	✓	✓

insurance industry.

The proposed methodology is justified by its capability to manage the complexities associated with insurance data, including varying data distributions and outliers. By combining advanced statistical techniques with optimization algorithms, our approach guarantees robust model performance and increases the interpretability of results. Furthermore, employing data-driven methodologies aligns with contemporary trends in the insurance industry, where data analytics increasingly influences strategic decisions and business outcomes.

In conclusion, the proposed methodology signifies a considerable advancement in insurance premium pricing, offering insurers a practical and scientifically sound approach to overcoming the challenges presented by heterogeneous data. By highlighting the research gap, emphasizing our contribution, and justifying our methodology, this study sets the foundation for future research initiatives and underscores the significance of data-driven approaches in shaping the future of insurance analytics.

3 Methodology

This section presents the research method. To design the research model, data collection was followed by preprocessing and sampling. Once the data was prepared, several data mining techniques were applied based on the type and condition of the data. In the next stage, the reduction of input variables was analyzed, as it was crucial to determine how this reduction impacted the model's accuracy, according to the research data. The genetic algorithm significantly enhanced the model's accuracy in the subsequent stage. The final steps of the model incorporated fuzzy logic for logical estimation.

3.1 Sampling Method and Data Analysis

To address the issue of data heterogeneity in our study, we utilized a stratified sampling approach. Stratified sampling involves dividing the population into distinct subgroups, or strata, based on relevant characteristics such as demographic variables or risk profiles. This method aims to ensure that each subgroup is adequately represented in the sample, thereby minimizing potential bias and enhancing the generalizability of the results.

Exploratory Data Analysis:

Before implementing the stratified sampling method, we conducted a comprehensive exploratory data analysis (EDA) to illustrate the presence of data heterogeneity. EDA included examining the distribution of key variables, such as age, income levels, and health status across the dataset. This analysis revealed notable variations in these variables, confirming the presence of heterogeneity. For instance, the variance in income levels and health statuses among different demographic groups indicated that

simple random sampling might not adequately capture the population's diversity.

Justification for Stratified Sampling :

Given the findings from the EDA, stratified sampling was deemed suitable for addressing the identified heterogeneity. By partitioning the dataset into strata based on the identified variables, we ensured that each subgroup was internally homogeneous while maintaining variation among different strata. This method allows us to capture the full spectrum of data characteristics and enhance the precision of our model's predictions. Each stratum was sampled in proportion to its size within the population, ensuring that every observation had an equal chance of being included in the sample, thus improving the representativeness of the data.

Modelling Procedure and Analysis:

In addition to stratified sampling, this section outlines the modelling procedure and overall analysis process. After stratifying the data, we applied various modelling techniques to evaluate their performance. The results were then analyzed to determine the effectiveness of stratified sampling in managing data heterogeneity. This comprehensive approach ensures that the modelling process is informed by a well-represented sample, leading to more accurate and reliable predictions. The structure and model designed to address the current problem are illustrated in Figure 1. To resolve the issue, the steps outlined in Figure 1 were implemented.

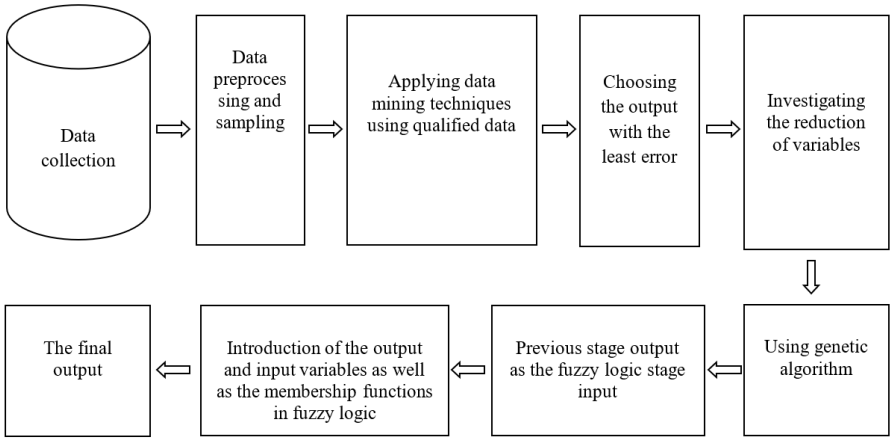

Figure 1: The model of the problem

As illustrated in Figure 1, after collecting the data, steps 1 to 5 focus on estimating household expenses, constituting the first phase of the research. Our objective is to estimate the insurance premium. Consequently, steps 6 and 7, which incorporate fuzzy logic, are included in the model. The following sections will review and analyze the steps depicted in Figure 1.

In the insurance industry, attention is often paid to the damage incurred when set-

ting insurance premiums; hence, the variables that directly influence damage should be taken into account during data collection. Therefore, salary figures and the number of dependents were considered input variables in this study. To further enhance the robustness of the methodological framework, the study adopts a multi-layered approach to data preprocessing and model optimization. After ensuring a representative dataset through stratified sampling, the preprocessing stage addresses potential data inconsistencies and prepares the data for advanced modelling. Specifically, feature selection is performed to eliminate redundant or irrelevant variables, thus reducing dimensionality and improving the predictive power of subsequent models. This step is crucial in preventing the models from overfitting or being misled by noise within the dataset.

Normalization is applied to standardize the data range, particularly because variables such as salary and health status vary significantly in scale. This process is essential for the proper functioning of distance-based algorithms, like Support Vector Machines (SVM), ensuring that the optimization process does not favour variables with larger scales. Additionally, shuffling is used to minimize any sequence effects in the data, thus preventing the model training process from being affected by the order in which data is presented.

During the model selection phase, a combination of nonlinear data mining (DM) techniques is utilized to capture the complex relationships within the heterogeneous data. Techniques such as quadratic and cubic SVMs, along with Gaussian Process Regression, are selected for their capability to model nonlinear interactions. These methods are particularly well-suited for the insurance sector, where risk factors can interact in intricate ways that linear models may overlook. The model's performance is thoroughly evaluated, and the parameters of the neural network, such as weights and biases, are fine-tuned using Genetic Algorithms (GA) to achieve the lowest Mean Squared Error (MSE), thus enhancing the model's predictive accuracy.

Finally, to convert the model predictions into actionable insurance premiums, fuzzy logic is employed. This approach facilitates the integration of expert knowledge into the decision-making process, offering a flexible mechanism for adjusting premiums based on nuanced, continuous input variables. The use of membership functions within the fuzzy logic framework ensures that the output is not only precise but also interpretable, adhering to industry standards and practices. This comprehensive methodological approach, which integrates advanced statistical and machine learning techniques, reinforces the reliability and applicability of the study's findings.

3.2 Data Set

The research data utilized in this study were extracted from the internal records of a large insurance organization, covering the fiscal year 2019. The dataset includes detailed information on employees' monthly salaries and the number of dependents, categorized by subordination types, which served as the primary input variables.

Specifically, numerical values were assigned to different subordination types based on the relationships of the dependents to the insured employees. For instance, the children of an employee, classified as Subordinate 1, are their primary dependents, while Subordinate 2 refers to the employee's spouse (if the employee is female). Subordinate 3 encompasses other dependents, such as the employee's parents. This classification is crucial for differentiating between the various categories of dependents and evaluating their impact on household expenses.

The output variable, the total cost incurred by the entire household on behalf of the organization, encompasses expenses related to claims and other associated costs. This cost data, alongside the input variables, was gathered from a diverse sample of insured employees, resulting in a dataset comprising 2,224 individual records. The substantial size of the dataset ensures that the analysis captures a wide range of scenarios, thereby enhancing the robustness of the study's findings.

The following sections present key concepts related to data mining (DM), data heterogeneity, neural networks, genetic algorithms (GA), fuzzy logic, and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) to provide a clearer understanding of the topic.

Table 2: Variables of the insured individuals

Input variable	**Input variable**			**Output variable**
	Number of subtypes			
Salary	Subordinate 1	Subordinate 2	Subordinate 3	**Household expenses**

3.3 Data Mining

In short, the information hidden in a set of data is obtained by analysis of definite DM approaches, among which the most important ones are recommending and suggesting, identifying incorrect data, risk management, analysis of performance, customer segmentation, targeted advertisement, and prediction. Generally, DM is oriented by four methods and tools such as clustering, classification, participation, and regression rules. Classification is often done through algorithms along with observers. That is, learning is based on some input and output data in such a way that part of the data serves to train the system and quantify the characteristics of the method, and then it is used for the input whose output is not available. Some DM algorithms that typically use observer learning are Decision Tree, Naive Bayes, KNN, NN, SVM, GA and Rugged Set Approach.

In the context of this research, Data Mining (DM) techniques play a critical role in analyzing complex and heterogeneous data, particularly in the insurance industry. Given the large volumes of data generated from various sources such as employee records, household expenses, and insurance claims, traditional methods are often insufficient for extracting meaningful patterns and making accurate predictions. Therefore,

advanced DM techniques were employed to address these challenges effectively.

The primary DM methods used in this study include Support Vector Machines (SVMs), Neural Networks (NNs), and Genetic Algorithms (GA). These methods were selected based on their proven effectiveness in handling non-linear and high-dimensional data, as well as their ability to adapt to the complexities inherent in heterogeneous datasets. SVMs, for example, are particularly well-suited for classification tasks and have been extensively used in the insurance domain to predict customer behaviour and risk factors.

Neural Networks were also employed due to their capability to model complex relationships between input variables and outputs. In this study, NNs were used to predict household expenses based on variables such as employee salary, number of dependents, and previous claims history. The network's architecture was carefully designed to minimize the Mean Squared Error (MSE) during training, ensuring high accuracy in predictions. To further enhance the performance of the NNs, Genetic Algorithms were applied to optimize the network's weights and biases. This combination of techniques allowed for a more refined and accurate prediction model, capable of handling the intricacies of the data.

In summary, the application of advanced DM techniques in this study was critical for addressing the challenges posed by heterogeneous data in the insurance industry. By combining NNs, GA, and Fuzzy Logic, the study was able to produce a robust model that can be used for more accurate prediction of household expenses and determination of insurance premiums. This methodological approach not only aligns with best practices in data mining but also represents a significant advancement in the application of these techniques to complex real-world problems.

3.4 Data Heterogeneity

According to Frost (2019), Heterogeneity is determined as a difference between elements comprising a whole. By existing heterogeneity, the characteristics under study become diverse. The parts of the whole are various, not the same. It is a key concept in statistics and science. The opposite of homogeneous is heterogeneous. Heterogeneity is a vital concept in statistics appearing in different contexts, thus altering its definition. Heterogeneity can represent differences within individual specimens, samples, and experimental outcomes in a meta-analysis. Heterogeneous data are any data with high variability of data formats and types. Possibly, they are low quality and ambiguous owing to high data redundancy, missing values, and untruthfulness. Hence, heterogeneous data are difficult to integrate to satisfy the demands of business information.

Data heterogeneity typically refers to differences in the structure, type, and format of data, such as variations between structured, semi-structured, and unstructured data. However, within the scope of this research, we focused on the heterogeneity in terms of variability and diversity within structured data, particularly how the same set of

input variables can lead to different output values due to complex underlying factors not captured by the selected variables. Variables such as the type of disease in a family or the availability of various medical services in the area of residence of people can change the output answer.

Table 3: Example of Study Data.

		Input variable		Output variable
Salary	Subordinate1	Subordinate2	Subordinate3	Household Expenses
15,630,000	1	0	0	366,020
15,630,000	0	0	0	1,490,050
15,630,000	0	0	0	446,000
15,630,000	0	0	0	827,000
15,630,000	0	0	0	127,000
15,630,000	0	0	2	3,763,640
15,630,000	0	0	0	3,938,400

In Table 3, an example of the data of the current study is presented. As you can see, a large number of records have different output values for the same inputs. And this is a challenge to solve.

3.5 Neural Networks

A neural network consists of hidden, input, and output units. Inputs enter the network from the left, activate the hidden units in the middle, and exit through the output units on the right. Learning in neural networks through feedback is referred to as a Backpropagation neural network. This involves comparing the network's output with the expected output and using the difference between these two results to adjust the weights of the connections between network units. By applying this algorithm repeatedly, we undergo two stages. Initially, when artificial neural networks are formed, the weight values are assigned randomly. After weighing, we progress through the Feed Forward process, where the input data are multiplied by the weights and subsequently accumulated with bias. Ultimately, this phase yields an output value that may diverge from the actual output. Once the network detects the error, we transition to the second part of this stage, where the network recalculates weights and biases. This adjustment continues as long as the predicted output remains less than the actual output. In designing neural networks, data should be divided into three segments: training, testing, and evaluation, with testing and evaluation data comprising a small percentage of the total data. Additionally, the choice of training and transfer functions is crucial in neural network design, as these parameters significantly aid in reducing errors during training, testing, and evaluation.

3.6 Genetic Algorithms (GA)

The process of natural selection is effectively simulated by Genetic Algorithms (GA), demonstrating how species adapt to environmental changes to thrive and reproduce

across generations. This principle, encapsulated in the notion of "survival of the fittest," is employed to solve complex problems. Each generation is composed of individuals, each representing a point in the search space and a potential solution, resembling a chromosome. GA draws its foundation from the genetic structure and behaviour of chromosomes within a population. It serves as a method for addressing constrained and unconstrained optimization problems, rooted in the principles of natural selection that underpin biological evolution. Consequently, GA proves beneficial for various optimization challenges that are ill-suited for standard algorithms, particularly when the objective function is random, discontinuous, or highly nonlinear. GA is also adept at solving mixed integer programming problems, where some variables are restricted to integer values. In this study, GA was employed to fine-tune the weights of the neural network. To enhance the efficiency of the neural network and minimize error, we train the network and then utilize the GA to feed in the optimal weights. These weights are adjusted through the fitness function using coefficients ranging from zero to one. We then redesign the neural network, inputting the optimized weights extracted from GA before training the network with these modified weights. GA is emphasized as a distinct component of our methodology due to its unique role in optimizing the neural network architecture. It is a heuristic optimization technique inspired by natural selection and evolution, repeatedly generating a population of candidate solutions—termed chromosomes—and evaluating their fitness based on a predefined objective function. Through iterative processes such as selection, crossover, and mutation, GA evolves the population. In our study's context, GA optimizes the architecture of a neural network model, specifically the selection of network topology, activation functions, and learning parameters. By encoding potential network configurations as chromosomes and evaluating their performance on a validation dataset, GA identifies the optimal parameter set that maximizes the model's predictive accuracy. By presenting GA as a core component of our methodology, we underscore its role as a vital optimization tool that complements the decision-making techniques employed in our study. Through the continual refinement of the neural network architecture, GA enhances the robustness and generalizability of the prediction model, ultimately improving its performance in generating superior estimates.

3.7 Fuzzy Logic

Fuzzy logic analyzes entities on a continuum from zero to one, rather than a binary scale of zero and one. This approach accommodates an infinite number of nodes or degrees of membership, closely aligning with human reasoning and real-world situations. Since the 1970s, fuzzy logic has been a staple in system control. The system's accuracy and effectiveness are defined by control rules that dictate the relationships between inputs and outputs. Consequently, fuzzy logic has gained extensive use in data classification and clustering. Some studies have shown that fuzzy neural models yield superior results in specific cases compared to traditional models.

3.7.1 Fuzzy Logic in This Study

In this study, a fuzzy model was implemented to determine the premium paid by the insured. The data output served as the input for the fuzzy model. A membership function was defined for each variable, essentially measuring its degree of membership. The relevant parameters are as follows:
a) Average household expense
b) Average premium paid

It is noteworthy that for parameters a and b, the membership function is defined as 1. After establishing the membership functions and variables in the software, it was essential to formulate the rules. In the fuzzy model, the relationship between the input and output membership functions is positive, indicating that higher household expenses correlate with increased premiums paid by the insured.

4 Results

This section analyzes the results obtained after designing the model and applying the data. The study aims to evaluate the effectiveness of the proposed premium determination strategy, which is based on household costs compared to the conventional salary-based approach. First, we examined and analyzed the strategy of determining insurance premiums based on salary, followed by the strategy of determining insurance premiums based on household expenses, which is the model proposed in this study. In the proposed model, we initially test the decision-making techniques appropriate for the model, and after selecting the best technique, we work on reducing the error. We evaluate the sampling and assess whether modifications to the number of input variables improve our results. Additionally, to further reduce errors, we apply the genetic algorithm to validate our findings. At the end of the initial stage of the proposed model, specifically regarding household expenses, we compare the results with other methods to gauge the efficiency of our model. Subsequently, we can evaluate the sensitivity of the outputs to the input variables. Ultimately, we conduct the premium calculation test and examine the comparative effectiveness of the two premium determination strategies.

4.1 Salary-Based Premium

One of the key variables in determining insurance premiums is the salary of the insured individuals. Insurance companies often prioritize individuals who have higher salaries but lower associated costs in to mitigate their risks. Consequently, this section delves into the analysis of insurance premiums based on salary levels. To gain insights, we categorize the data into six salary-based groups. By investigating the relationship between salary levels and incurred costs, we aim to identify factors that could potentially reduce expenses for insurance companies.

4.2 Analyzing Salary-Based Premium

The salary variable plays a significant role in determining insurance premiums. Insurance companies generally seek individuals with high salaries or income levels who present low costs. As a result, several articles indicate that insurance companies rank the insured based on their income levels. The dataset comprises 2,224 entries, with an average salary of 40,040,743, a minimum salary of approximately 10 million, and a maximum salary of approximately 70 million. To ensure no category has zero frequency, we divided the data into six categories based on salary, allowing for category adjustments as required. For optimal categorization, we established a category width of 10 million. Reducing costs in the insurance industry can effectively decrease associated risks. Thus, insurance companies actively seek factors that could help lower these costs.

Table 4: Income classification and related household expenses.

Class	# of Cases	Sum of Expenses	Mean Household Expenses
Salary ≥ 60,000,000	73	155,037,080	2,123,796
50,000,000 ≤ Salary < 60,000,000	221	553,506,200	2,504,553
40,000,000 ≤ Salary < 50,000,000	856	1,658,962,074	1,938,040
30,000,000 ≤ Salary < 40,000,000	575	912,571,318	1,587,081
20,000,000 ≤ Salary < 30,000,000	412	792,521,160	1,923,595
Salary < 20,000,000	87	73,873,130	849,116.4

In Table 4, salaries are categorized into six levels. The lowest average household expenses are associated with individuals earning the least. Those with salaries below 20,000,000 Rials (the monetary unit) have an average of 849,116 in household expenses. The results from the table indicate that higher-income individuals impose more costs on the insurance organization. Since insurance premiums are calculated as a percentage of a person's salary, those with higher earnings pay higher premiums. These individuals argue that because they contribute more to insurance premiums, they should incur more costs for the organization. Consequently, they do not embrace a culture of prevention. Their higher premium payments psychologically lead them to burden the insurance organization even further. Therefore, basing insurance premiums on a percentage of salary is not ideal for insurance companies, as the premiums collected are comparatively small relative to the expenses incurred. The data indicates a direct relationship between salary and the expenses imposed on the insurance organization.

4.3 Comparison with the Proposed Model

Before analyzing salary-based premiums in detail, it's important to establish a baseline by comparing the traditional approach to the designed model. In this subsection, we outline the experimental setup and specific research questions addressed in each experiment. Clearly defining the objectives of each experiment ensures a cohesive and clear analysis. This step is vital for a comprehensive understanding of the proposed model's performance compared to conventional methods.

4.4 Experimental Design

To effectively address the research questions, we have designed a series of experiments to assess the performance of the proposed premium determination model under various scenarios. These experiments are carefully constructed to analyze how different factors—such as household costs, income levels, and risk profiles—impact the accuracy of premium estimations. Variables like age and type of disease informed our subordinate factor analysis. Referrals of higher-income individuals to medical institutions contributed to our understanding of the salary factor. As mentioned in previous sections, exploratory data analysis (EDA) was used to review the distribution of key variables, including age, income level, and health status, within the dataset. By systematically analyzing the results, we can draw significant conclusions about the efficacy and applicability of the proposed model in real-world insurance contexts.

4.5 Analyzing Expenditure-Based Premiums

Data mining (DM) and fuzzy logic techniques determine insurance premiums based on household costs. DM methods are leveraged in the insurance industry to predict and mitigate risk. However, using predictive methods such as neural networks (NN) and adaptive neuro-fuzzy inference systems (ANFIS) alone does not fulfil our requirements. Our objective is to estimate premiums fairly. This study also investigates the effects of insured salaries and the number and nature of subordination on household expenses and premium determination using DM techniques. Assessing the current research methodology lays the groundwork for further discussion regarding the estimated model and the research questions. Neural networks are one of the various DM techniques, and a comparison of these techniques is presented in Section 4.2.1.

4.5.1 Neural Network Results

Determining premiums for insured individuals is to manage the costs associated with health insurance. To achieve this, non-linear DM methods were applied to the problem data, including quadratic support vector machines (SVM), complex trees, cubic SVM, exponential Gaussian process regression, fine Gaussian SVM, and NN (as shown in Table 5). The data for the current model is non-deterministic and non-linear; thus, DM techniques were selected to accommodate the data's characteristics.

Table 5: Evaluation of data mining techniques.

Method	Neural Network	Complex Tree	Fine Gaussian SVM	Quadratic SVM	Cubic SVM	Exponential GPR
MSE Value	0.0002	0.0064	0.0049	0.0064	6.5025	0.0049

We identified the most effective technique based on the lowest Mean Squared Error (MSE). Our evaluation included all six techniques tested. Complex tree and quadratic Support Vector Machine (SVM) methods yielded identical MSE values of 0.0064. Similarly, the fine Gaussian SVM and exponential Gaussian process regression also achieved

an MSE of 0.0049. However, the cubic SVM technique exhibited a considerably high error, indicating its unsuitability for the model. According to Table 5, we selected the Neural Network (NN) method for the data analysis. The evaluation process halts when there is no improvement after six iterations, marking the optimal strategy for training the neural network. In our NN framework, 15% of the data is allocated for testing and another 15% for model evaluation at each stage. The best training algorithm established is the Levenberg-Marquardt (LM) method, with the number of hidden layers configured to 10. Despite the selection of the neural network over other data mining techniques, as illustrated in Figure 2, a noticeable error still persists within the network. This discrepancy can be attributed to the heterogeneity of the data, R=0.38.

The neural network regression diagram indicates that the training and test data learning processes are underperforming. When the neural network's learning is insufficient, accurate predictions cannot be achieved.

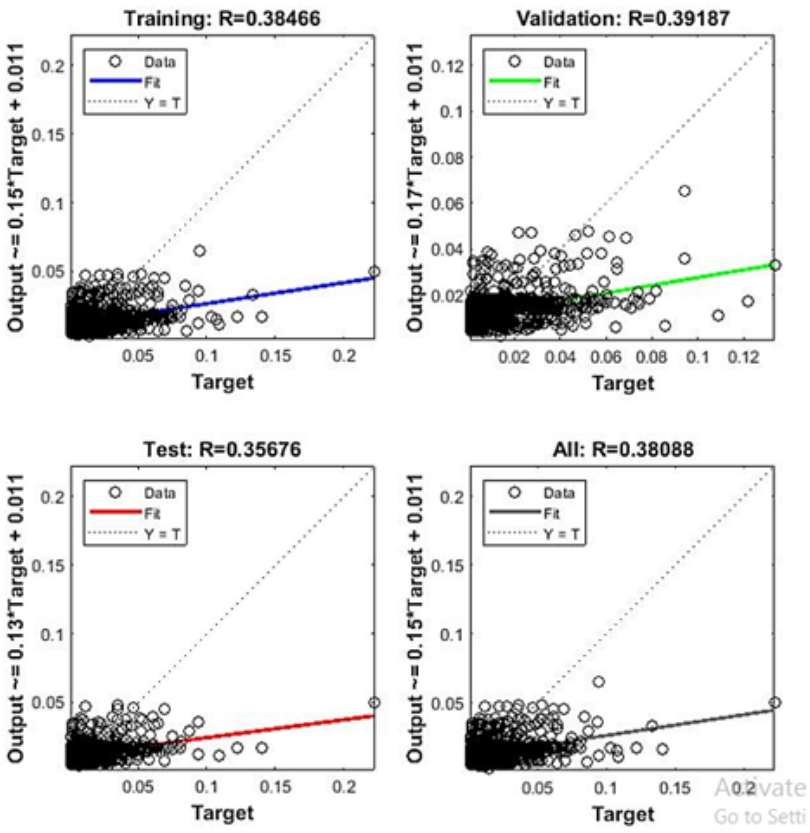

Figure 2: Neural network regression diagram (initial).

4.5.2 Utilizing Sampling to Reduce Error

Stratified sampling is a form of purposive sampling used in research, which involves selecting members that are both abundant and diverse. In this case, approximately 10% of the data was chosen as a sample. Since data shuffling has already been completed, the selected sample conforms to the conditions of the entire dataset.

As illustrated in Figure 3, sampling has a profound impact on reducing error, with the correlation coefficient increasing to R=0.85. Specifically, the correlation for the training data rose from R=0.38 to R=0.86, while the test data improved from R=0.35 to R=0.85, enhancing our optimism for accurate predictions.

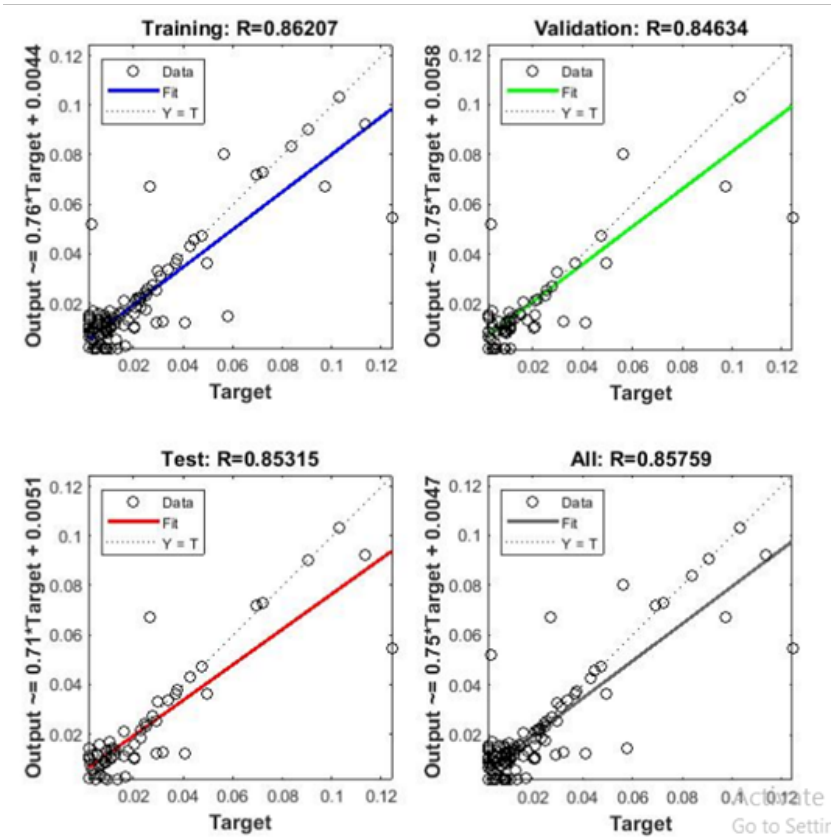

Figure 3: Neural network regression diagram after applying sampling.

4.5.3 Sensitivity Analysis

Our ongoing goal is to minimize errors further. A pertinent question that arose was whether removing one or more input variables would lead to a decrease in error. If we decide to eliminate one or two subordinate variables, we will retain the variable that exerts the most influence on the output while discarding the less impactful ones. This

phase involves evaluating the reduction of input variables as a strategy to decrease errors. In this study, we employed a neural network approach instead of solely relying on the correlation coefficient method.

Table 6: Sensitivity analysis

Initial value	Increase in sub 3	Increase in sub 2	Increase in sub 1	Error (sub 3)	Error (sub 2)	Error (sub 1)
2,639,810	2,548,948	2,089,379	2,636,785	90,862	550,431	3,025
2,111,697	2,360,590	2,027,394	2,136,875	248,893	84,303	25,178
1,625,148	1,857,669	1,781,149	1,836,048	232,521	156,001	210,900
2,156,424	2,651,682	2,740,422	2,221,390	495,258	583,998	64,966

The reduction of subordinate variables is illustrated in Table 6. It is evident that subordinate variable 2 exhibits a greater dependence on the output due to its higher error rate. Consequently, subordinate variables 1 and 3 have been eliminated.

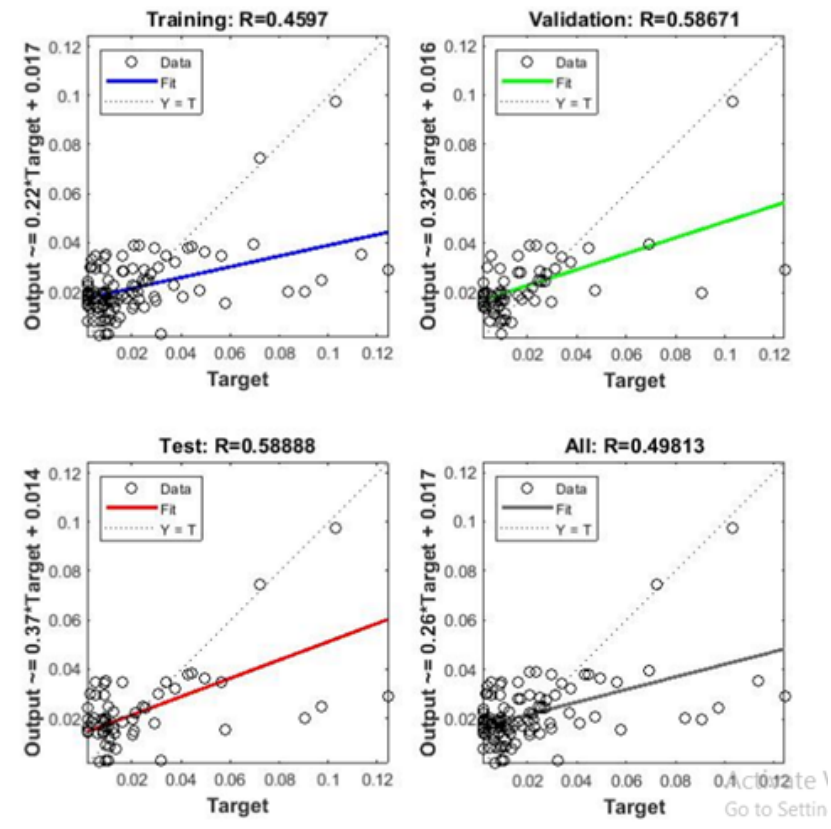

Figure 4: Regression after reducing input variables.

Unfortunately, as illustrated in Figure 4, the network error has risen once more. By removing two input variables, we observed an increase in the similarity among the remaining input variables. Therefore, we will continue with the previous variables.

It seems logical that when there is a significant amount of similar data, removing a variable can lead to a greater number of similar records, while adding a variable may reduce the number of similar records, ultimately enhancing the effectiveness of data mining techniques.

4.5.4 The Impact of Genetic Algorithms

One valuable characteristic of neural networks is their ability to improve weights. To enhance the efficiency of the neural network and minimize errors, we first design the network and then train it. We input the network's weights into the minimum function of the Genetic Algorithm (GA). In this process, the weights are modified by multiplying them with coefficients ranging from zero to one within the fitness function. Afterwards, we redesign the neural network and input the weights derived from GA into the network before retraining it with these adjusted weights. As depicted in Figure 5, the performance and efficiency of the network have significantly improved, reaching an R-value of 0.91.

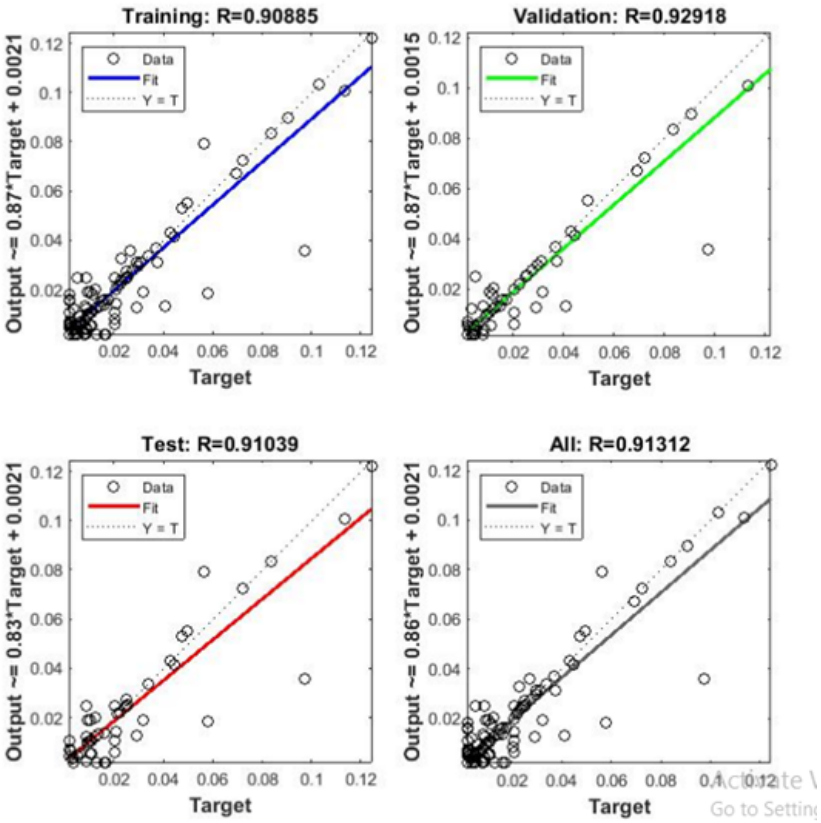

Figure 5: Neural network regression after GA optimization.

The correlation coefficient for the training data improved from R=0.86 to R=0.90,

while the test data showed an increase from R=0.85 to R=0.91. Although there remains a possibility of errors in prediction calculations due to the data's heterogeneity, an overall value of 91 is deemed acceptable.

4.5.5 Evaluating the First Stage

In our study, we selected fourteen specific data points to evaluate and compare the performance of our current model with the Adaptive Neuro-Fuzzy Inference System (ANFIS). These data points were chosen for their representativeness of various scenarios within the dataset, ensuring a comprehensive assessment of the model's performance across different conditions. The selection encompassed a range of values for input variables such as salary and number of subordinates, along with their corresponding real household expenses, thereby reflecting typical cases encountered in the insurance domain. This strategic approach allowed us to rigorously test the model's robustness and accuracy in predicting household expenses.

Table 7 presents a detailed comparison between actual conditions, predictions made by our current model, and those generated by ANFIS. It includes a side-by-side look at the household expenses and the predicted values from both models for the selected data points. The evaluation metrics employed to gauge model performance are the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), chosen for their capacity to indicate the model's accuracy and error distribution. The MAE assesses the average magnitude of prediction errors, while the RMSE evaluates the prediction error with greater emphasis on larger errors.

Table 8 offers further insights through sensitivity analysis of the input variables, examining how variations in salary and the number of subordinates affect the predicted household expenses. Sensitivity analysis is crucial for understanding individual variables' influence on model outputs and validating the model's responsiveness to input changes. The data illustrated in Table 8 reflect the impact of fluctuating salaries and subordinate counts on predicted expenses, which contributes to assessing the model's reliability in diverse scenarios. Figure 6 enhances this analysis by visualizing the sensitivity of the input variables. It displays three series showing changes in the input variables and their corresponding effects on output, aiding comprehension of the model's sensitivity to input data variations and providing a graphical representation of the findings detailed in Table 8.

In summary, the added explanations, detailed evaluation metrics, sensitivity analysis, and visual representations, yield a thorough assessment of the model's performance. This comprehensive approach ensures clarity when comparing our current model to ANFIS, underscoring the advantages and robustness of our proposed method in predicting household expenses.

Table 7: Evaluating the first stage

Monthly salary	Sub 1	Sub 2	Sub 3	Real Condition	Current model	ANFIS
46,699,445	3	1	0	1,801,000	1,946,856	29,789,000
47,132,886	3	1	0	2,670,700	2,216,189	30,348,000
36,321,750	0	1	0	319,000	294,749	2,461,400
56,290,441	4	1	1	5,163,000	10,111,366	156,640,000
15,630,000	0	1	0	765,000	769,191	885,270
35,130,413	0	0	0	635,000	347,791	931,000
47,427,000	4	0	0	635,000	561,698	2,660,000
49,507,829	0	0	0	1,477,940	2,774,632	1,310,000
66,613,300	4	0	3	765,000	765,207	1,090,000
22,276,695	0	2	4	2,409,000	2,409,887	2,410,000
52,350,443	4	0	0	1,146,000	515,752	54,822,000
40,024,836	4	0	0	882,430	5,001,245	52,444,848
36,867,385	3	0	0	127,000	248,633	11,034,000
29,334,713	0	0	0	598,050	1,724,625	1,079,700

The error metrics, specifically Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) demonstrate a substantial improvement in our model compared to the ANFIS method. According to the data presented in Table 7, our current method outperforms the competition. The MAE for the ANFIS model stands at 2.35 e+07, while the MAE for our model is just 9.45 e+05.

Table 8: Sensitivity analysis of input variables

Output related to Table 7	Output related to a 10,000,000 rial salary increase	Output related to the increase in subordinate number 2
1,946,856	17,370,407	17,370,407
2,216,189	17,370,407	17,370,407
294,748/8	248,633	17,135,964
10,111,366	17,103,061	17,370,407
769,191	16,455,818	17,370,393

We can analyze the sensitivity of the input variables. As illustrated in Figure 6, three distinct series of changes were applied to the inputs.

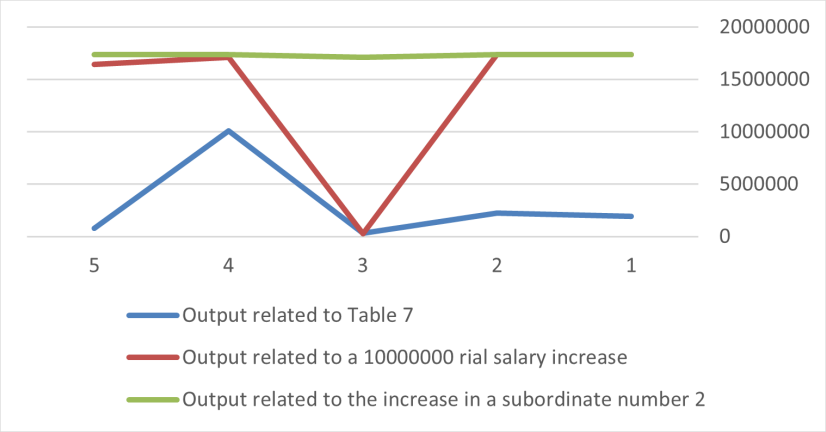

Figure 6: Sensitivity analysis of the model's first-stage output.

Series 1 presents output related to the data in Table 7. Series 2 pertains to a 10,000,000 rial increase in the salary variable. It's important to note that Row 3 has no subordinates, and the salary increase does not influence the output. Series 3 examines the impact of increasing the number of subordinate 2, which significantly affects output more than subordinates 1 and 3 (as detailed in Section 4.2.3). By enhancing subordinate 2, there is a more pronounced change in output. This emphasizes the necessity of focusing on subordinates, and future research should delve into this aspect. Additionally, factors such as personal health and environmental conditions are essential considerations that warrant exploration in subsequent studies. As a result, the output can now be predicted with greater accuracy. In the initial stage, output variables are utilized as inputs for fuzzy logic.

4.5.6 Calculation of Insurance Premiums

A fuzzy model was implemented to calculate the premiums paid by the insured. To assess the overall model and evaluate the effects of inputs on outputs at each stage, five samples were randomly defined for each of the aforementioned stages. The primary objective of this project is to foster a culture of prevention among the organization's clients, specifically the insured, and to reduce overall costs for the organization in the long term. Table 9 illustrates several examples of how insured individuals pay their premiums. Notably, the insurance premium calculation is not based on 7% of salary, but rather on household expenses.

Table 9: The final model format to determine the premium

Monthly salary	Sub 1	Sub 2	Sub 3	First-stage output (household cost)	Combined model output (premium)
46,699,445	3	0	0	1,946,856	2.50E+06
47,132,886	3	0	0	2,216,189	2.50E+06
36,321,750	0	0	0	294,748/8	5.71E+05
56,290,441	4	0	1	10,111,366	4.56E+06
15,630,000	0	0	0	769,191	6.60E+05

The table above indicates that individuals with higher salaries typically incur greater costs for insurance organizations. As a result, they are required to pay higher premiums.

Table 9 is replicated in Table 10, which includes an additional column titled "Receive Premium Based on 7% of Salary." A comparison of these two tables reveals that determining insurance premiums based on household expenses aligns more closely with the principles of social justice. In Row 4 of Table 10, we see an individual with a high income whose premium based on household expenses exceeds the premium calculated as a percentage of their salary. Conversely, Row 5 features a low-income individual whose premium based on household expenses is lower than that based on their salary percentage. This trend continues across the other rows in the table. Given the previously established link between household expenses and income, this model proves beneficial for insurance organizations in determining fair and equitable

premiums.

Table 10: Comparing two strategies

Salary	Household expenses strategy	7% of salary strategy
46,699,445	2.50E+06	3.27E+06
47,132,886	2.50E+06	3.30E+06
36,321,750	5.71E+05	2.54E+06
56,290,441	4.56E+06	3.94E+06
15,630,000	6.60E+05	1.09E+06
35,130,413	5.81E+05	2.46E+06
47,427,000	6.27E+05	3.32E+06
49,507,829	3.73E+06	3.47E+06
66,613,300	6.81E+05	4.66E+06
22,276,695	2.60E+06	1.56E+06
52,350,443	6.16E+05	3.66E+06
40,024,836	3.72E+06	2.80E+06
36,867,385	5.64E+05	2.58E+06
29,334,713	2.50E+06	2.05E+06

Table 11: Comparison with recent studies

Monthly salary	Sub 1	Sub 2	Sub 3	Real Condition	Current model	Özgur & Yolcu	Jones & Swati
46,699,445	3	1	0	1,801,000	1,946,856	2,624,976	1,660,222
47,132,886	3	1	0	2,670,700	2,216,189	2,624,869	3,480,578
36,321,750	0	1	0	319,000	294,749	1,629,579	994,037
56,290,441	4	1	1	5,163,000	10,111,366	3,802,120	3,926,635
15,630,000	0	1	0	765,000	769,191	1,646,733	582,824
35,130,413	0	0	0	635,000	347,791	1,213,821	454,000
47,427,000	4	0	0	635,000	561,698	2,414,252	5,579,650
49,507,829	0	0	0	1,477,940	2,774,632	966,709	1,381,771
66,613,300	4	0	3	765,000	765,207	4,876,900	4,518,444
22,276,695	0	2	4	2,409,000	2,409,887	2,181,282	5,366,627
52,350,443	4	0	0	1,146,000	515,752	2,397,081	1,283,485
40,024,836	4	0	0	882,430	5,001,245	2,362,322	938,086
36,867,385	3	0	0	127,000	248,633	2,334,124	1,089,016
29,334,713	0	0	0	598,050	1,724,625	1,255,331	700,104

We re-evaluated the validity of the current study by comparing it with recent research in the field. Table 11 examines household medical expenses for the studies conducted by Özgur and Yolcu (2023) and Jones and Swati (2023). An overview of these studies can be found in Section 2.5. Using the sample data presented in the table, we calculated the error of the results by comparing them to the actual values. The Mean Absolute Error (MAE) for Özgur and Yolcu's study is 1.23 e+06, while Jones and Swati's study has an MAE of 1.16 e+06. In contrast, the MAE for the current study is 9.45 e+05, indicating its superiority.

5 Discussion

The findings suggest that the developed model effectively addresses the research questions posed. Overall, this research incorporates the household expenses of the insured

into their insurance premiums, distinguishing it from other studies in the field, particularly those focused on health insurance. Furthermore, it emphasizes the role of insured individuals in cost management. The acceptance of the model's results can be attributed to the reduction of data heterogeneity. Without addressing data heterogeneity, achieving reliable results would have been impossible, which was one of the study's objectives. Factors such as stratified sampling, the appropriate application of data mining techniques, the implementation of genetic algorithms, and the use of fuzzy logic contributed to diminishing data heterogeneity and enhancing model efficiency. The designed model can be easily utilized within insurance systems. Each year, data necessary for the model, including membership function values, is defined, allowing users to calculate insurance premiums by entering relevant input variables. A simple program linked to the algorithm can facilitate all these processes.

An analysis of the literature review table demonstrates the alignment of this study's findings with previous research across several dimensions. Firstly, the study's focus on product sales and customer needs resonates with past research by Wanke & Barros (2016), Abdi et al. (2017), and Boodhun & Jayabalan (2018), which emphasizes the significance of customer-centric strategies in the insurance industry. By integrating data mining (DM) techniques with advanced modelling approaches, this research provides a sophisticated analysis of customer behaviour and its effects on sales and premium strategies, showcasing a clear progression in meeting the evolving needs of the insurance market, particularly regarding customer satisfaction and retention.

Secondly, this study examines the issue of damage to insurers, relevant to previous research by Cigsar & Unul (2019) and Yin et al. (2021), who investigated factors contributing to insurance claims and damages using DM techniques to predict and mitigate risks. The current research builds on this foundation by incorporating not only traditional DM techniques but also Genetic Algorithms (GA), enhancing the model's ability to manage data heterogeneity, an essential factor in accurately assessing and predicting damage-related risks. While consistent with prior findings, this study expands the boundaries by addressing more intricate data scenarios and delivering robust and comprehensive solutions.

Additionally, this research contributes to efforts aimed at reducing fraud in the insurance sector, a critical issue explored by Amini & Abdi (2018), Reddy et al. (2019), and Sharma (2021). These studies investigated fraudulent activities through various DM techniques, underscoring the necessity for accurate data analysis. The present study strengthens these initiatives by employing advanced DM methods combined with fuzzy logic to refine fraud detection and prevention. The integration of fuzzy logic offers a nuanced decision-making approach, which is especially useful in fraud detection where precision is crucial. This novel method not only aligns with past studies but also enhances their methodologies, providing an effective toolkit for combatting fraud in the insurance industry.

Finally, the study's emphasis on data heterogeneity and its influence on model performance denotes a significant contribution that aligns with and extends the work of

researchers like Owens et al. (2022), Goodarzi & Janat Babaei (2016), Zhao et al. (2021), and Wilson et al. (2024). The innovative combination of DM techniques with GA introduced in this study marks a significant methodological advancement, ensuring more accurate and reliable outcomes. By addressing the complexities of heterogeneous data, this research offers invaluable insights essential for developing effective premium strategies, thereby making a considerable contribution to the existing body of knowledge.

5.1 Managerial Insights

The implications of this research extend beyond academia, providing strategic guidance for managers and practitioners in the insurance industry. Here, we explore the nuanced applications and managerial considerations arising from our findings:

Strategic Premium Calculation

Our study highlights the necessity of innovative premium calculation strategies aligned with fairness and equity principles. By incorporating household expenses into premium assessments, managers can cultivate a more equitable insurance landscape that ensures premiums reflect individual risk profiles and financial capabilities.

Cost Optimization and Revenue Enhancement

Utilizing our proposed model allows managers to optimize organizational costs while boosting revenue streams. By accurately assessing insurance premiums based on household expenses, insurers can mitigate financial risks related to mispricing and maintain a balance between revenue generation and customer satisfaction.

Customer-Centric Service Delivery

At the core of our findings lies the principle of customer-centricity, where insurers prioritize their clients' needs and preferences. By leveraging insights from our model, managers can customize insurance products and services to meet the diverse demands of their customer base, fostering long-term loyalty and trust.

Risk Management Amidst Heterogeneity

In an era marked by data heterogeneity, our research equips managers with a robust toolkit for effective risk management. By addressing the complexities of heterogeneous data through our model, insurers can make informed decisions, identify emerging trends, and proactively mitigate risks, thereby safeguarding the financial health and stability of their organizations.

Utilizing Model Outputs for Customer Valuation

Beyond premium calculation, our model's outputs can serve as invaluable resources for customer valuation and segmentation. By leveraging the differences between household expense-based premiums and traditional salary-based premiums, managers can gain insights into customer lifetime value (LTV) and tailor marketing and retention

strategies accordingly.

Enhancing Organizational Agility and Competitiveness

Ultimately, adopting our model fosters organizational agility and competitiveness in a dynamic insurance landscape. By embracing data-driven insights and innovative methodologies, insurers can position themselves as industry leaders, ready to adapt to evolving market trends and seize emerging opportunities.

6 Conclusion

If insurance premiums are set reasonably, substantial savings can be realized, benefiting both the insured individuals and society at large. However, premiums are often calculated inaccurately, overlooking key variables such as household expenses and the number of dependents. This study introduces novelty by addressing these overlooked factors in social health insurance, highlighting their impact on insurance premiums. The challenge posed by data heterogeneity was minimized through advanced techniques; a neural network evaluated at 0.0002 was ultimately chosen for data analysis. Sampling and Genetic Algorithms (GA) significantly enhanced the neural network's performance, resulting in superior prediction outcomes compared to the ANFIS method. Yet, this stage of evaluation did not address service pricing; therefore, a fuzzy logic model was employed as the most suitable tool to align all initial outputs with the problem outputs. The model's acceptability is attributed to the reduction of data heterogeneity. Without this reduction, achieving effective results would have been impossible, which was one of the primary objectives of this study.

The comparison between insurance premium strategies based on salary versus those based on household expenses revealed that the latter yields better outcomes for insurance organizations. The model developed in this research has the potential to minimize out-of-pocket expenditures, foster financial stability in the healthcare system, encourage a preventative culture, assess customer value, mitigate insurance risks, enhance public health, and ensure appropriate premiums for dependents.

Subsequent research can explore all the aforementioned aspects, including the impact of this study on organizational cost reduction, increased income, and enhanced satisfaction among insured individuals. A number of future studies have been suggested:

- Determining the premium of subordinates according to the type of subordination with advanced computational techniques.
- The cultural and economic impact of the premium based on the cost in the social insurance sector on the insurer and the insured.
- Identification of fraud in the insurance industry through artificial intelligence (in the case of heterogeneous data).

Data Availability Statements

The data that support the findings of this study are available from the Health Insurance Organization of Iran but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Data are however available from the corresponding author upon reasonable request.

References

Abdi, F., Khalili-Damghani, K., and Abolmakarem, S. (2017). Solving Customer Insurance Coverage Sales Plan Problem Using a Multi-Stage Data Mining Approach. *Kybernetes*, **47**(1), 2–19.

Boodhun, N., and Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms.

Brofer, A., Rezaian, A., and Shokoohyar, S. (2017). Identification of Customer Behavior Pattern in Life Insurance and Capital Formation Using Data Mining. *Management Research in Iran*, **20**(4), 65–94.

Folland, S., Goodman, A., and Stano, M. (2016). *The Economics of Health and Health Care*. Routledge. https://doi.org/10.4324/9781315510736.

Frost, J. (2019). Heterogeneity. *statisticsbyjim.com/basics/heterogeneity*.

Ghuse, N., Pawar, P., and Potgantwar, A. (2017). An Improved Approach for Fraud Detection in Health Insurance Using Data Mining Techniques. *International Journal of Scientific Research in Network Security and Communication*, **5**(5).

Goel, S., and Chaudhary, A. (2024). Prediction of Health Insurance Price using Machine Learning Algorithms. *INDIACom*, **2024**. DOI:10.23919/INDIACom61295.2024.10498661.

Goodarzi, A., and Janat Babaei, S. (2016). Evaluation of Decision Tree Algorithms, Naive Bayes and Logistic Regression in Detection of Car Insurance Frauds. *Insurance Research Quarterly*, **1**(2), 61–80.

Jones, K. I., and Swati, S. (2023). The Implementation of Machine Learning in the Insurance Industry With Big Data Analytics. *International Journal of Data Informatics and Intelligent Computing*, **2**(2), 21–38.

Kalra, M., Lal, N., and Qamar, S. (2018). K-Mean Clustering Algorithm for Mining Heterogeneous Data. *Information and Communication Technology for Sustainable Development*. DOI:10.1007/978-981-10-3920-1-7.

Kalra, H., Singh, R., and Kumar, T. S. (2022). Fraud Claims Detection in Insurance Using Machine Learning. *Journal of Pharmaceutical Negative Results*. https://doi.org/10.47750/pnr.2022.13.S03.053.

Kumar Dubey, A., Kumar Dubey, A. N., Agarwal, V., and Khandagre, Y. (2012). Knowledge discovery with a subset–superset approach for Mining Heterogeneous Data. *CSI Sixth International Conference on Software Engineering (CONSEG)*. DOI:10.1109/CONSEG.2012.6349495.

Özgur, B., and Yolcu, U. (2023). Prediction of the Premium Production of Insurance Companies Operating in Turkey Using Artificial Neural Networks. *Turkish Journal of Forecasting*. https://doi.org/10.34110/forecasting.1223653.

Panda, S., Purkayastha, B., Das, D., Manomita, C., and Saroj, B. (2022). Health Insurance Cost Prediction Using Regression Models. *COM-IT-CON*, **2022**. DOI:10.1109/COM-IT-CON54601.2022.9850653.

Pantelous, A., and Passalidou, E. (2013). Optimal premium pricing policy in a competitive insurance market environment. *Annals of Actuarial Science*, **7**(2), 175–191.

Patil, M. S., Sanika, K., and Sanjana, K. (2024). Medical Insurance Premium Prediction with Machine Learning. *International Journal of Innovations in Engineering Research and Technology*. https://doi.org/10.26662/ijiert.v11i5.pp5-12.

Rezaei Navaei, S., and Koosha, H. (2016). Applying Data Mining Techniques for Customer Churn Prediction in Insurance Industry. *International Journal of Industrial Engineering & Production Management*, **27**(4), 635–653.

Rose, F. (2013). *Marine Insurance: Law and Practice*. Routledge.

Salama, M., Abdelkader, H., and Abdelwahab, A. (2022). A novel ensemble approach for heterogeneous data with active learning. *International Journal of Engineering Business Management*. https://doi.org/10.1177/18479790221082605.

Timothy, J., Layton, A., Randall, P., Thomas, G., and Van Kleefd, R. (2017). Measuring efficiency of health plan payment systems in managed competition health insurance markets. *Journal of Health Economics*, **56**, 237–255.

Voto, T., and Ngepah, N. (2025). Out-of-Pocket Health Expenditure in Sub-Saharan Africa. *Economies*. https://doi.org/10.3390/economies13050119.

Wang, T. Ch., and Liaw, R. T. (2020). Multifactorial Genetic Fuzzy Data Mining for Building Membership Functions. *IEEE Congress on Evolutionary Computation (CEC)*. DOI:10.1109/CEC48606.2020.9185900.

Wanke, P., and Barros, C. (2016). Efficiency drivers in Brazilian insurance. *Economic Modelling*, **53**, 8–22.

Wilson, A. A., Nehme, A., Dhyani, A., and Mahbub, K. (2024). A Comparison of GLM with Machine Learning Approaches for Predicting Loss Cost in Motor Insurance. *Risks*. https://doi.org/10.3390/risks12040062.

Yan, C. H., Li, Y., Liu, W., Li, M., Chen, J., and Wang, L. (2019). An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification.