Global Journal of Economic and Finance Research

e-ISSN: 3050-5348
p-ISSN: 3050-533X

Vol. 02(09): 886-912, September 2025

Home Page: https://gjefr.com/

Actuarial Science and Artificial Intelligence: Optimization and Deep Learning for Risk Modeling in Morocco 2022

Hamza Boucheta[1], Marouane Karim[2], Agouram El Mahjoub[3], Noura Moummadi[4], Lamia Mourtadi[5], Faris Asmaa[6], Elhachloufi Mostafa[7]

[1,2,3,4,5]Master's student in Participatory Finance Engineering and Artificial Intelligence, Faculty of Legal, Economic and Social Sciences – Ain Sebaa, Hassan II University, Casablanca, Morocco

[6]Laboratory of Applied Modeling for Economics and Management, Faculty of Legal, Economic and Social Sciences – Ain Sebaa, Hassan II University, Casablanca, Morocco

[7]Department of Statistics and Applied Mathematics for Economics and Management, Hassan II University, Casablanca, Morocco

KEYWORDS: automobile insurance, premium pricing, generalized linear models (GLM), credibility theory, risk distribution.

Corresponding Author:
Marouane Karim

Publication Date: 22 September-2025
DOI: 10.55677/GJEFR/17-2025-Vol02E9

ABSTRACT

This study explores the integration of actuarial science and artificial intelligence through a hybrid framework combining optimization techniques and deep learning to enhance actuarial risk modeling. Using a Moroccan case study, we leverage linear programming for financial decision-making and arti- ficial neural networks (ANN) for predicting and analyzing actuarial risks under uncertainty.

The methodology is applied to real-world data from the Moroccan insurance sector, demonstrating its practical relevance for improving long-term financial planning and risk evaluation.

Experimental results show that the ANN model achieves a normalized Mean Absolute Error (MAE) of 0.625, outperforming traditional Generalized Linear Models (GLMs) by approximately 15% in pre- dictive accuracy. Risk assessment metrics further reveal improvements, with the 95% Value at Risk (VaR) and Tail Value at Risk (TVaR) reduced by 9.4% and 13.9%, respectively, compared to Poisson and Negative Binomial GLMs. These gains highlight the ANN's capacity to capture complex nonlinear relationships and heteroscedasticity in insurance claim data.

Moreover, linear programming integration enables optimized premium pricing and reserve allocation under multiple regulatory and economic constraints.

This hybrid approach fosters a more robust and adaptive actuarial process, equipping insurers to bet- ter manage uncertainty and extreme risk scenarios.

Overall, this work modernizes actuarial practices by introducing intelligent, data-driven models that enhance predictive performance and strategic decision-making, contributing to sustainable financial stability in emerging markets such as Morocco.

1. INTRODUCTION

In the context of modern financial systems, insurance serves as a fundamental mechanism for trans- ferring and managing risk. Individuals and institutions are continuously exposed to uncertain events that can result in significant financial losses.

As economic activities expand and risks become more complex, the need for accurate and scientific approaches to insurance pricing and risk allocation has become increasingly pressing Bühlmann, 1967; Denuit et al., 2007.

Insurance companies must determine premiums that not only ensure financial sustainability but also reflect the true risk profile of each policyholder.

The traditional methods of premium calculation often based on historical averages or heuristic assump- tions are no longer sufficient to

meet the demands of a data-driven and highly competitive insurance market Anisetti et al., 2021; Antonio and Valdez, 2012; Klugman et al., 2012.

This growing complexity has led to the emergence of mathematical methodologies that combine ele- ments of probability theory, actuarial science, and optimization to construct pricing models capable of handling uncertainty, heterogeneity among insureds, and regulatory constraints Bühlmann and Gisler, 2005; Charpentier, 2020.

Among the most significant of these methods are probabilistic risk models, such as the Poisson and Cramér-Lundberg models, which provide frameworks for estimating claim frequencies, aggre- gate losses, and ruin probabilities Klugman, 1992; Zehnwirth, 2002. In parallel, risk measures such as Value-at-Risk (VaR) and Tail Value-at-Risk (T-VaR) have become essential tools for quantifying potential extreme losses Klugman et al., 2012.

In addition, modern developments in computational mathematics have enabled the use of optimiza- tion algorithms such as linear and nonlinear programming, genetic algorithms, and ensemble machine learning methods like XGBoost to derive optimal premium values under various economic and actu- arial constraints Frees, 2010; Wüthrich, 2018.

The objective of this research is to analyze and implement these mathematical methodologies within the context of insurance, with a particular focus on their application in pricing and risk allocation . Special attention is given to the Moroccan automobile insurance market, where challenges related to data availability, fairness, and risk variability remain prevalent Rouyan and Amrani, 2020.

The study proposes a modeling framework based on real or simulated insurance data, where key vari- ables influencing premium determination are extracted, analyzed, and integrated into a quantitative pricing system.

A computational algorithm developed in Python is used to automate the estimation of fair premiums while managing exposure to high-risk segments through an Adjustment Coefficient.

By combining theoretical foundations with practical implementation, this work contributes to the advancement of actuarial and mathematical practices in insurance.

It aims to offer robust tools for insurers to determine risk-based prices, allocate risk efficiently, and support regulatory compliance, profitability, and customer trust.

2. LITERATURE REVIEW

The insurance sector is a fundamental pillar in enhancing financial stability and protecting individuals and institutions from risks.

It also plays a significant investment role by mobilizing financial resources and directing them towards productive projects.

Pricing policies are among the most critical aspects influencing the dynamics of the insurance market, reflecting a company's ability to balance the attractiveness of offers and ensure profitable sustainabil- ity.

Many studies have relied on fundamental theories and concepts in insurance, such as Credibility Theory (Bühlmann, 1967) & (Herzog, 1999) & (Mahler and Dean, 1999) & Generalized Linear Models (GLM) (McCullagh and Nelder, 1989) & (de Jong and Heller, 2008); (Ohlsson and Johansson, 2010), which provide accurate premium estimates based on driver and vehicle characteristics.

Recent literature highlights the necessity of employing precise mathematical models for determining insurance premiums that reflect the actual risk degree and ensure fairness among insured parties. Among the most used models are collective risk models, Poisson models, and the Cramér-Lundberg model, which analyzes the probability of ruin and links premiums to risk levels and contract nature, as detailed in (Antonio and Valdez, 2012) & (Zehnwirth, 2002) & (Taylor and Ashe, 1983) & (Gong et al., 2018).

A prominent study by (El Attar et al., 2019) proposed a novel mathematical approach for optimal reinsurance parameters based on a dual criterion combining the minimization of ruin probability and maximization of the insurer's technical utility.

The study used the Cramér-Lundberg model and solved constrained equations via Genetic Algorithms, providing a precise method for optimal insurance pricing under uncertain risks (Chen and Guestrin, 2016), (Charpentier, 2020), (Liu and Shih, 2011) .

What distinguishes this study is its integration of mathematical theory and applied optimization using computational intelligence techniques, serving as an important reference for updating pricing models in various branches, including automobile insurance.

The same concept can be adapted to estimate premiums based on driver and vehicle characteristics and driving behavior while considering the company's financial safety margin.

Although some field studies exist in the Moroccan context, such as (Rouyan and Amrani, 2020) on policyholder switching motives, most rely on traditional statistical tools (logistic regression) without addressing advanced mathematical models or quantitative analysis tools for precise pricing.

In the Maghreb region, two applied studies from Algeria are noteworthy: one forecasting loss rates in automobile insurance at Boulfekhar Company using Box-Jenkins time series models, concluding that these rates are predictable and useful for improving reinsurance and pricing decisions; and another by Zerman analyzing risk management procedures at the Boulfekhar Mila branch, emphasizing the importance of a precise field approach to accident classification and evaluation in risk management. A clear gap exists in Moroccan literature regarding the use of modern mathematical models for pricing automobile insurance contracts, such as **Credibility Theory** (Credibility Theory) and the Bühlmann and Bühlmann-Straub models (Bühlmann, 1967), which integrate individual and collective experience in pricing.

Additionally, GLM models and risk measures like VaR and T-VaR provide effective tools for analyzing and forecasting extreme losses

under varying uncertainty levels (Denuit et al., 2007).

The use of mathematical optimization techniques such as **non-linear programming** and **genetic algorithms** enables the determination of optimal pricing under specific regulatory or technical con- straints.

These methods have been applied in reinsurance studies and can be adapted to automobile insurance (Chapman et al., 2000) , (Shearer, 2000, (Hastie et al., 2009),(Janssens et al., 2006).

Based on the above, this study aims to fill this knowledge gap by building an applied mathematical model for pricing automobile insurance contracts in the Moroccan market.

It relies on realistic technical and demographic data and employs advanced Python-based software tools for data analysis and fair, precise premium allocation.

3. METHODOLOGY

This study relies on a quantitative mathematical approach to determine optimal pricing for car insur- ance contracts, through the use of advanced optimization models, actuarial and probabilistic models, supported by specialized software tools.

The methodology adopted is inspired by the approach in El Attar et al., 2019, which combines max- imizing underwriting profit and minimizing the probability of ruin, using genetic algorithms as an effective tool for nonlinear optimization.

The theoretical foundations of the adopted methodology rely primarily on actuarial models, which form the basis for premium estimation based on risk quantification.

3.1 Actuarial Models

The study is based on actuarial pricing principles grounded in credibility theory and risk measurement:

- **Expected value:**

Premium = E[X] + Loading

- **Actuarial fairness principle:**

Each contract is priced according to the individual risk degree based on technical and demo- graphic characteristics.

In addition to actuarial pricing principles, probabilistic models play a central role in modeling claim frequency and severity, allowing for more nuanced risk assessments.

3.2 Probabilistic and Statistical Models

Used models include:

- **Frequency distributions:** Poisson, Binomial, Negative Binomial
- **Severity distributions:** Gamma, Pareto, Exponential, Lognormal
- **Survival analysis and stochastic processes:** Markov chains and renewal processes.

Frequency Distributions of Claims:

Table 1: Frequency Distributions of Claims

Assumptions	Mathematical Formula	Distribution
$\lambda > 0$	$P(N = k) = \frac{\lambda^k e^{-\lambda}}{k!}$	Poisson
$n \in \mathbb{N},\ 0 < p < 1$	$P(N = k) = \binom{n}{k}(1-p)^{n-k}p^k$	Binomial
$r > 0,\ 0 < p < 1$	$P(N = k) = \binom{r+k-1}{k}(1-p)^r p^k$	Negative Binomial

The Poisson distribution is typically used under the assumption of a constant event rate and equidispersion (mean = variance), whereas the negative binomial distribution offers more flexibility by accommodating overdispersion—common in insurance data.

Severity Distributions for Modeling Loss Amounts:

Table 2: Severity Distributions for Modeling Loss Amounts

Distribution	Density Function	Parameters	Description
Exponential	$f(x) = \lambda e^{-\lambda x}$	$\lambda > 0$	Small losses
Gamma	$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$	$\alpha, \lambda > 0$	Flexible and asymmetric
Pareto	$f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$	$\alpha > 0, x_m > 0$	Large losses, heavy tails
Lognormal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$	$\mu, \sigma > 0$	High dispersion

The choice of severity distribution depends on the shape of the empirical loss data.

For instance, Pareto and lognormal distributions are well suited to capture heavy-tailed behavior and high variability in claims, which are common in insurance settings with catastrophic risk potential.

3.3 Artificial Neural Network (ANN) Architecture

In this study, a supervised learning approach based on an Artificial Neural Network (ANN) was de- veloped to predict insurance premiums.

The model was designed as a feedforward neural network consisting of one input layer, one hidden layer, and one output layer.

The input layer receives seven normalized and encoded features considered to be significant predic- tors of premium pricing: sexe_encoded, age, vehicle_age, power, license_age, passengers, and fuel_type_encoded.

These variables were selected based on actuarial relevance and prior statistical analysis.

The hidden layer contains four neurons, each activated using the Rectified Linear Unit (ReLU) acti- vation function, which introduces non-linearity and helps the model learn complex patterns.

The output layer consists of a single neuron with a linear activation function to produce a continuous output corresponding to the predicted premium value.

The network was compiled using the Mean Squared Error (MSE) as the loss function, optimized with the Adam optimizer.

The model was trained over 100 epochs with a batch size of 16, using early stopping to prevent over- fitting.

The dataset was split into training and testing sets with an 80/20 ratio to evaluate the model's gen- eralization capacity.

Table 3: Model Summary of the Artificial Neural Network

Parameter	Description
Model Type	Feedforward Neural Network (Dense ANN)
Input Features	7 (sexe_encoded, age, vehicle_age, power, license_age, passengers, fuel_type_encoded)
Hidden Layers	1 hidden layer with 4 neurons (ReLU activation)
Output Layer	1 neuron (Linear activation)
Loss Function	Mean Squared Error (MSE)
Optimizer	Adam
Epochs	100
Batch Size	16
Validation Method	80% training / 20% testing split

This architecture was selected for its simplicity and efficiency, providing a balance between pre- dictive power and interpretability.

The diagram below illustrates the information flow across the ANN model layers.

The following algorithm outlines the training pipeline of the ANN model used for premium prediction:

Algorithm 1 Neural Network Training for Insurance Premium Prediction

Input: Dataset with features and target premium
Output: Trained neural network model and prediction results
Load dataset from Excel file
Select input features: sexe_encoded, age, vehicle_age, power, license_age, passengers, fuel_type_encoded
Extract target variable: prime_estimee
Normalize input features using StandardScaler
Optionally normalize the target variable
Split dataset into training (80%) and testing (20%) sets
Define neural network model:
Input layer with 7 features
Hidden dense layer with 4 neurons, ReLU activation
Output dense layer with 1 neuron, linear activation
Compile model using Adam optimizer and Mean Squared Error loss
Set early stopping on validation loss (patience = 10)
Train model on training data with:
Epochs = 100
Batch size = 16
Validation split = 10%
Evaluate model on test set and compute MAE
Predict premiums and inverse-transform normalized outputs
Compare predicted and actual premiums

The training algorithm described above outlines the step-by-step construction and optimization of the artificial neural network

designed for premium prediction.

This architecture was chosen to balance model complexity and performance, leveraging a single hidden layer with a modest number of neurons to avoid overfitting while capturing non-linear relationships in the input variables.

The following figure provides a visual representation of the implemented network, highlighting the input features, hidden layers, and the output node corresponding to the estimated insurance premium.

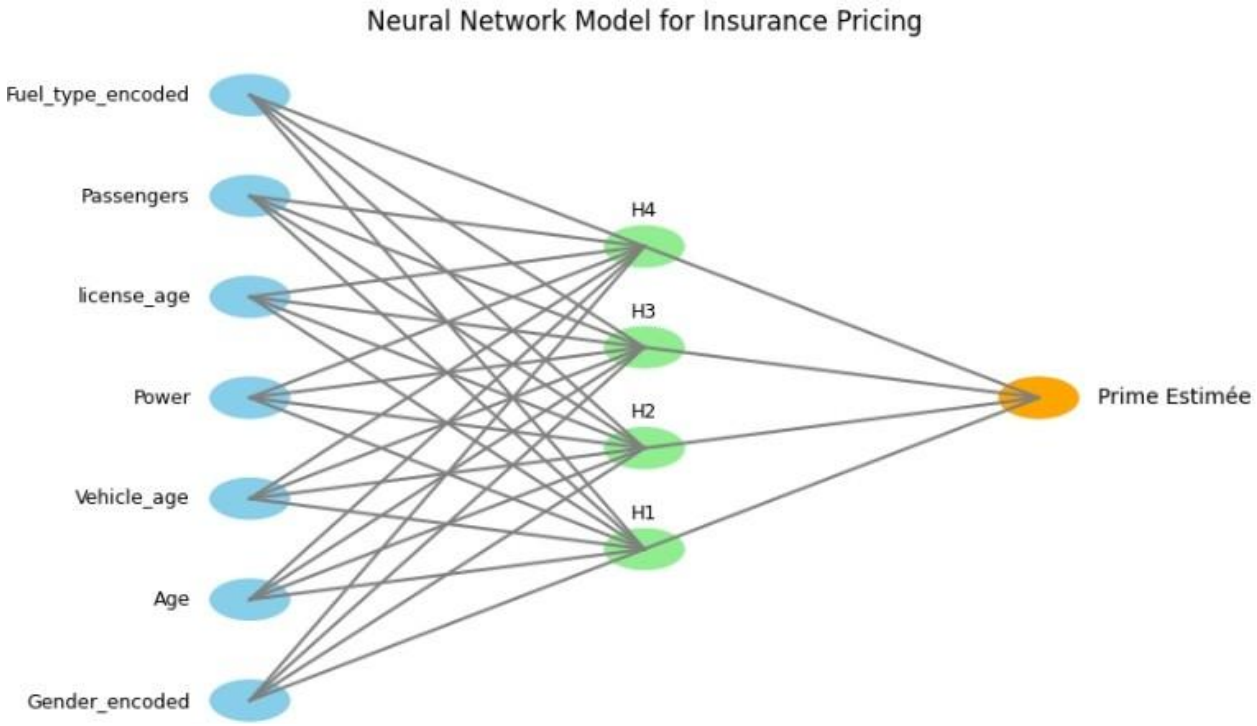

Figure 1: Architecture of the Artificial Neural Network used for premium prediction

3.4 Analytical Tools

To implement the actuarial and statistical models developed in this study, two main analytical environments were used: Python and IBM SPSS Modeler.

Each of these tools offers complementary advantages, contributing to an efficient, rigorous, and replicable data mining process.

The main tools and libraries employed include:

- **IBM SPSS Modeler**: used primarily for data exploration, preprocessing, and the construction of visual workflows.

Its drag-and-drop interface facilitates rapid prototyping and model validation without extensive coding, especially for classification and regression tasks.

- **Python (Pandas, NumPy)**: employed for deeper statistical analysis and custom data manipulation.

These libraries were essential in transforming the raw insurance dataset into a structured format ready for advanced modeling.

- **Scikit-learn**: for the implementation of machine learning algorithms and evaluation metrics, including those used to assess premium prediction models.
- **Matplotlib, Seaborn**: to generate detailed visualizations such as boxplots, heatmaps, and distribution plots that support the exploratory phase and aid in interpretation.

By combining the visual power of SPSS Modeler with the flexibility of Python, this project bene- fited from a dual approach that ensured both interpretability and control over model implementation.

3.5 Data Used

This study utilizes a set of real variables extracted from car insurance contracts within the Moroccan insurance market, as reported in prior research on insured behavior and insurer switching Rouyan and Amrani, 2020.

A mathematical model was developed to characterize the individual risk profile of each insured, based on their technical and demographic characteristics.

This model is implemented through programming algorithms in **Python**, enabling precise premium calculation and adjustment via a risk modification factor grounded in actuarial and statistical methods. Furthermore, the dataset was used to compute advanced risk measures, such as Value at Risk (VaR) and Tail Value at Risk (T-VaR).

The impact of applying a **Stop-Loss** reinsurance contract on loss distribution was simulated to assess risk mitigation strategies Cai and Tan, 2000; Gerber, 1972; Kaas et al., 2008; Tan and Weng, 2006; Wang et al., 1998.

This integration of empirical and simulated data allowed for thorough testing of the model's ef- fectiveness in a near-real environment, facilitating scenario analysis and optimization of reinsurance strategies without requiring actual contract execution.

3.6 Applying the CRISP-DM Methodology Using SPSS Modeler

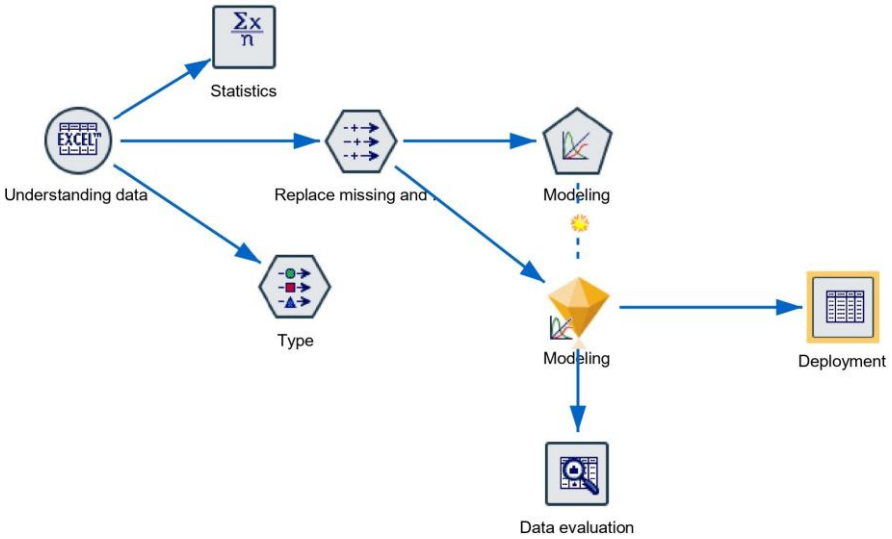

Figure 2: CRISP-DM (Cross Industry Standard Process for Data Mining)

The figure above illustrates the full data analysis workflow using **SPSS Modeler**, following the **CRISP-DM** methodology, which is the most widely adopted framework for data mining projects across different contexts.

This academic project systematically followed this methodology through a sequence of steps ensuring data quality and modeling effectiveness.

The process begins with the **Data Understanding** phase, where the database was loaded from an Excel file, representing the foundational step to explore the overall data structure and characteristics.

3.6.1 Software and Analytical Tools

This study leveraged a combination of specialized software tools and programming environments to support the various phases of the CRISP-DM methodology.

Each tool was selected based on its ability to effectively address specific stages of the data mining process.

SPSS Modeler was used as the primary platform for both data preparation and predictive model development.

Its intuitive graphical interface allowed for efficient implementation of data transformation, feature engineering, and modeling techniques, particularly Generalized Linear Models (GLMs), which were central to the pricing analysis.

In parallel, the Python programming language was employed to support the exploratory data analysis (EDA) and preprocessing tasks.

Python provided enhanced analytical flexibility through powerful libraries such as:

3.6.1.1 NumPy and Pandas for data manipulation and preparation,

3.6.1.2 Matplotlib and Seaborn for visualization and pattern discovery,

3.6.1.3 and Scikit-learn for auxiliary statistical operations and feature encoding.

Additionally, custom scripts were written to detect and correct outliers using robust methods like the Interquartile Range (IQR).

While SPSS Modeler handled the modeling phase, Python was instrumental in ensuring data quality and interpretability throughout the early stages.

This dual-tool approach ensured a rigorous and efficient workflow, balancing visual modeling capabil- ities with programmable analytical depth.

3.6.2 Understanding the data

After identifying the data source and selecting the appropriate analytical tools, it is essential to thoroughly explore the dataset. This phase, called understanding the data, provides crucial information about the structure, quality, and semantics of the data prior to modelling.

In this study, which focuses on estimating car insurance premiums, a comprehensive exploratory data analysis (EDA) was conducted to detect underlying patterns, relationships, and inconsistencies.

The dataset includes both numerical and categorical variables.
The numerical variables include the driver's age, the number of years since obtaining a driving licence, the age of the vehicle, the fiscal power, and the number of passengers.
The categorical variables correspond to the type of fuel used by the vehicle (diesel or gasoline) and the gender of the driver.
The dependent variable is the estimated insurance premium (ESTIMATED_PREMIUM).
In order to obtain preliminary insights, descriptive statistics such as the mean, median, standard deviation, skewness, and kurtosis were calculated for each numerical variable.
Visualisation techniques including histograms, box plots, and correlation heatmaps were employed to identify potential outliers, distribution asymmetries, and multicollinearity among features. Particular attention was paid to data quality issues such as missing values and inconsistent entries.
The results of these analyses informed subsequent data preprocessing steps, feature engineering, and the choice of modelling techniques.

Table 4: Sample of the Automobile Insurance Dataset Before Encoding

Gender	Age	Vehicle Age	Fuel Type	Power	License Age	Passengers
M	56	16	Diesel	9	38	5
M	42	5	Diesel	6	13	5
M	35	3	Gasoline	7	9	5
F	28	1	Diesel	5	6	5
M	60	10	Diesel	8	35	5
M	50	12	Gasoline	10	31	5
M	33	2	Diesel	6	10	5
M	46	8	Gasoline	9	27	5
M	22	9	Diesel	6	4	5
M	39	6	Gasoline	6	16	5
M	55	15	Diesel	7	34	5
M	61	10	Diesel	9	39	5
CO	31	3	Gasoline	6	9	5
F	45	5	Diesel	8	25	5
M	38	4	Gasoline	7	15	5
F	26	2	Diesel	5	6	5
M	52	10	Gasoline	10	29	5

Before proceeding with the detection and treatment of missing values and outliers, it is essential to ensure that all variables, particularly categorical ones, are properly prepared for numerical processing. Most statistical and machine learning algorithms require numeric inputs to function correctly.
Therefore, a categorical variable encoding step was performed to convert qualitative variables into discrete numeric representations. This transformation facilitates the subsequent data cleaning, exploratory analysis, and modeling pro- cesses.
The following section details the encoding procedures applied to the fuel type (Diesel or Gasoline) and Gender (driver category) variables, including illustrative examples before and after encoding.

3.6.3 Data Preparation: Encoding of Categorical Variables

The original dataset contains two qualitative categorical variables: the fuel type of the vehicle, referred to as combustion, which takes the textual values (diesel) and (Gasoline), and the driver's category, referred to as Gender, with values (Company), (Men), and (Woman).
These variables cannot be directly used in most statistical or machine learning algorithms that require numeric inputs.
To address this limitation, an encoding step was performed to convert both categorical variables into discrete numeric variables. The chosen encoding schemes are as follows:
Diesel → 1, Gasoline → 2
Company → 1, Men → 2, Woman → 3
This transformation preserves the qualitative information while making the data compatible with modeling algorithms.
Table 5 shows a sample of the dataset before encoding, where the categorical variables appear as text.

Table 5: Sample of the Automobile Insurance Dataset before Encoding

Age	Vehicle_Age	Fuel Type	Power	License_Age	Passengers	Gendre
56.0	16.0	diesel	9.0	38.0	5	Company
42.0	5.0	diesel	6.0	13.0	5	Men
35.0	3.0	essence	7.0	9.0	5	Woman
28.0	1.0	diesel	5.0	6.0	5	Woman
60.0	10.0	diesel	8.0	35.0	5	Men

After encoding, the new dataset contains numeric variables Fuel type_encoded and Gender_encoded.
Table 6 shows a sample after encoding.

Table 6: Sample of the Automobile Insurance Dataset after Encoding

Age	Vehicle_Age	Combustion_encoded	Power	License_Age	Passengers	Gendre_encoded
56.0	16.0	1	9.0	38.0	5	1
42.0	5.0	1	6.0	13.0	5	2
35.0	3.0	2	7.0	9.0	5	3
28.0	1.0	1	5.0	6.0	5	3
60.0	10.0	1	8.0	35.0	5	2

Following the encoding of categorical variables, the dataset was prepared for further quality as- sessment.
Specifically, a descriptive statistical analysis was conducted to explore the characteristics of the en- coded data, focusing on the distributions and variability of key input variables.
This step facilitated the identification of any data quality issues before modeling.
Subsequently, a thorough detection process was undertaken to identify missing values and outliers within the encoded dataset, as these issues can significantly affect model performance and reliability if left unaddressed.

3.6.1 Handling Missing Values and Outlier

Ensuring data quality and integrity prior to model development is a critical step in any data science project.
In this study, a comprehensive diagnostic was conducted to detect the presence of missing values and outliers, given their potential to adversely affect model accuracy and stability.

1.1 Detection of Missing Values

A customized algorithm was implemented within the SPSS Modeler environment to systematically examine potential missing or invalid entries.
This procedure involved inspecting zero or null values in essential variables such as age, power, vehicle_age, license_age, passengers, Gender_encoded, and Fuel type_encoded, among others.

Algorithm 2 Replace Zero Values with Medians

Require: Dataset with variables: age, vehicle_age, power, licence_age, passengers, Gender_encoded, Fuel type_encoded
Ensure: Replace zero values with corresponding medians
1: Compute: median_age → MEDIAN(age)
2: Compute: median_vage → MEDIAN(vehicle_age)
3: Compute: median_pow → MEDIAN(power)
4: Compute: median_lic → MEDIAN(licence_age)
5: Compute: median_pass → MEDIAN(passengers)
6: Compute: median_gender → MEDIAN(Gender_encoded)
7: Compute: median_Fuel → MEDIAN(Fuel type_encoded)
8: **for all** rows in the dataset **do**
9: **if** age = 0 **then** set age → median_age
10: **end if**
11: **if** vehicle_age = 0 **then** set vehicle_age → median_vage
12: **end if**
13: **if** power = 0 **then** set power → median_pow
14: **end if**
15: **if** licence_age = 0 **then** set licence_age → median_lic
16: **end if**

17: **if** passengers = 0 **then** set passengers → median_pass
18: **end if**
19: **if** Gender_encoded = 0 **then** set Gender → median_Gender_encoded
20: **end if**
21: **if** Fueltype = 0 **then** set Fueltype → median_Fuel
22: **end if**
23: **end for**
24: **return** Cleaned dataset

The results of this analysis, supported by descriptive statistics, confirmed that the dataset was completely free of missing values.
This finding reflects a high degree of completeness and consistency across the variables.

Table 7: Handling of Missing Values

#	Target Variable	Missing Values	First	Last	Valid	Treatment Method
1	age	0	1	122	122	SMEAN(age)
2	vehicle_age	0	1	122	122	SMEAN(vehicle_age)
3	power	0	1	122	122	SMEAN(power)
4	license_age	0	1	122	122	SMEAN(license_age)
5	passengers	0	1	122	122	SMEAN(passengers)
6	Gender_encoded	0	1	122	122	SMEAN(sexe)
7	Fuel Type_encoded	0	1	122	122	SMEAN(combustion)

1.2 Detection and Correction of Outliers

Ensuring the robustness and statistical validity of the dataset before modeling required a systematic procedure to detect and correct potential outliers.
These extreme observations, if left unaddressed, could distort model training and compromise predic- tive performance. A dedicated Python script was implemented to facilitate this process.
The approach involved two main steps:
1. Visual detection of outliers via **boxplots** for key numerical features.
2. Automated correction using an **IQR-based clipping algorithm**.

To visualize the distribution of values and identify potential outliers, a series of boxplots were gen- erated for the following numerical variables: age, vehicle_age, power, license_age, passengers, prime_estimee, as well as the encoded categorical variables Gender_encoded and Fuel type_encoded. Figure 3 summarizes the distributions.

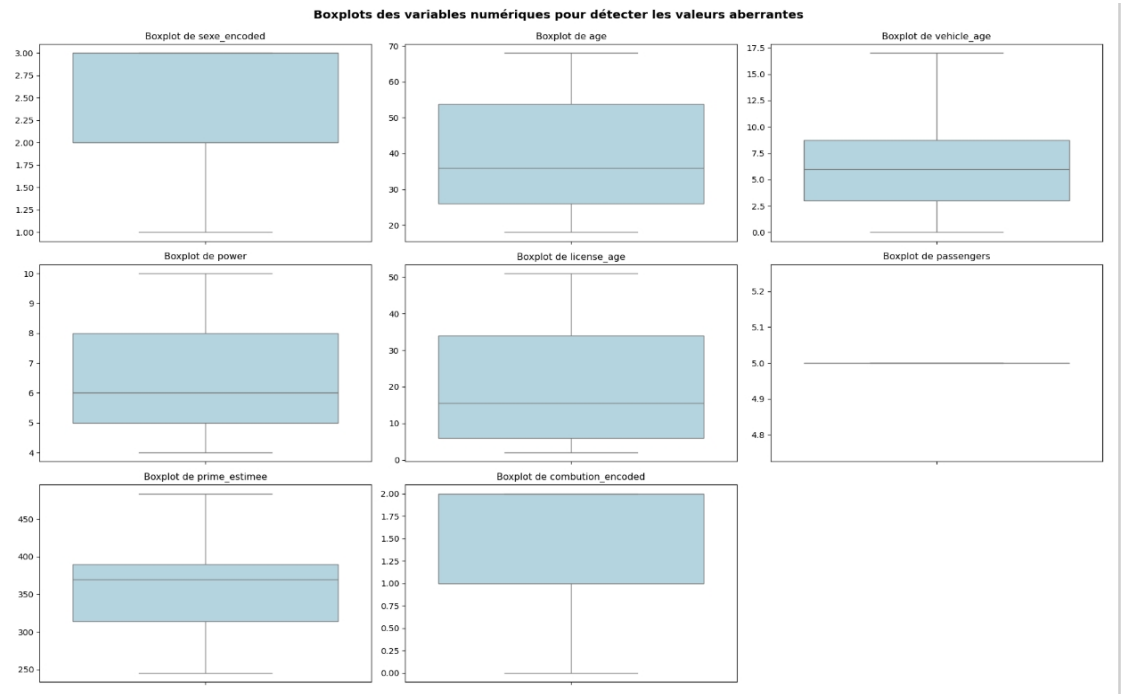

Figure 3: Boxplots of all numerical variables used to detect outliers.

Based on the visual analysis of Figure 3, several insights can be drawn. The variables age, vehicle_age, and license_age display moderately dispersed distributions with limited presence of outliers, indicating overall consistency in their values.
In contrast, both power and prime_estimee exhibit slight right-skewness, characterized by a few ob- servations that lie above the upper whisker and may be considered potential outliers.
The variable passengers remains constant across all records, which suggests that it provides minimal predictive value in its current form.
Lastly, the variables gender_encoded and fuel type_encoded, being encoded categorical features, naturally present discrete distributions, and what may appear as ”outliers” in their boxplots are ex- pected artifacts of their binary encoding rather than true anomalies.
To systematically correct potential outliers, an IQR-based clipping algorithm was developed, as shown in Algorithm 3.
This function iterates over all numerical variables and replaces extreme values with boundary limits defined by the interquartile range.

Algorithm 3 IQR-Based Outlier Clipping for Multiple Variables

Require: DataFrame df, List of numerical columns cols
Ensure: Return cleaned DataFrame df
1: **for each** column col in cols **do**
2: $Q1 \rightarrow$ 25th percentile of df[col] 3:
$Q3 \rightarrow$ 75th percentile of df[col] 4:
$IQR \rightarrow Q3 - Q1$
5: $low \rightarrow Q1 - 1.5 \times IQR$
6: $high \rightarrow Q3 + 1.5 \times IQR$
7: Clip all values in df[col] to the interval [low, high]
8: **end for**
9: **return** df

This method preserves the global structure and variability of the data while reducing the influence of extreme values.
Unlike deletion or winsorization, clipping maintains all observations in the dataset, thus avoiding data sparsity or bias introduced by removal.

3.6.2 Variable Type Assignment

To ensure the dataset was appropriately structured for analysis and modeling, a meticulous review of variable types was conducted.
The objective was to confirm that each variable was accurately classified as numerical, categorical, or logical, depending on its intrinsic properties and intended analytical role.
Using the ”Type” tool in SPSS Modeler, the dataset structure was standardized in accordance with best analytical practices.
This classification is critical, as it determines how each variable should be handled during preprocessing— whether it requires encoding, scaling, or exclusion.

Table 8: Description of the Input Variables

Field	Measure	Values	Missing	Check	Role
age	Continuous	[19.0, 68.0]	None	None	Input
vehicle_age	Continuous	[1.0, 17.0]	None	None	Input
power	Continuous	[5.0, 10.0]	None	None	Input
license_age	Continuous	[3.0, 51.0]	None	None	Input
Fuel Type_encoded	Binary	0, 1	None	Encoding applied	Input
passengers	Continuous	[5.0, 5.0]	None	Constant	Input
Gender_encoded	Binary	0, 1	None	Encoding applied	Input

Numerical variables, such as age, vehicle_age, and power, often require normalization or transfor- mation to ensure consistent scaling across features.
Categorical variables, including Fuel Type and Gender, were transformed into numerical format using binary encoding, thus enabling compatibility with machine learning algorithms. Finally, constant or low-variability variables, such as passengers, were reviewed for relevance, as their limited variation may provide little to no predictive value.

3.6.3 Descriptive Statistics

Descriptive statistical analysis was conducted to deepen understanding of the dataset structure and to study the basic statistical properties of the variables, especially those expected to directly influence premium pricing.

Table 9: Descriptive Statistics of Selected Variables Before SMOGN

Variable	Count	Mean	Median	Std	Min	Max
Gender_encoded	122	2.16	2.0	0.63	1	3
age	122	41.32	40.0	11.35	19	68
license_age	122	21.56	21.5	11.43	3	51
vehicle_age	122	5.89	5.5	3.52	1	17
power	122	6.74	6.0	1.40	5	10
passengers	122	5.00	5.0	0.00	5	5
Fuel Type_encoded	122	1.43	1.0	0.50	1	2
Estimated Premium	122	322.36	314.38	43.98	245.49	483.13

This analysis produced key descriptive statistics such as mean, median, standard deviation, mini- mum, and maximum values. These metrics are essential for identifying general trends, potential outliers, and data inconsistencies that may affect the robustness of predictive models.

The variables analyzed include demographic and technical characteristics directly affecting automobile insurance pricing.

The insured age ranges from 19 to 68 years, with a mean of 41.30 and standard deviation the proximity of 11.4, indicating a predominantly mature client base with moderate dispersion. The proximity of mean and median suggests a roughly symmetric distribution.

License age varies from 3 to 51 years, reflecting a broad range of driver experience. This diversity influences risk profiles, as younger license holders generally present higher accident probabilities.

Vehicle age ranges between 1 and 17 years with an average the proximity of 6.0 years, indicating a balanced mix of newer and older cars affecting premiums through depreciation and repair costs.

The power variable shows a narrow range from 5 to 10, with a mean the proximity of 6.70, suggesting most vehicles belong to a moderate engine power category, likely influenced by taxation or regulation. The number of passengers is constant at 5 for all observations, resulting in zero variance and no dis- criminative value for modeling. This might be a default or encoded placeholder.

Importantly, the variable Fuel type_encoded (1 for diesel, 2 for gasoline) indicates a nearly balanced split in fuel type, which is crucial as fuel type impacts risk exposure and maintenance costs, thus influencing premiums.

This descriptive phase plays a vital role in assessing data quality and highlighting variables needing transformation, review, or exclusion. Subsequently, variable type standardization was performed using SPSS Modeler's Type tool to correctly assign variables as numerical, categorical, or binary for proper modeling integration.

Its application led to a noticeable reduction in the extreme values of prime_estimee, particularly for gasoline-powered vehicles. Consequently, the cleaned dataset retained both statistical integrity and predictive richness.

3.6.4 Data Augmentation Using SMOGN

After conducting the descriptive statistical analysis, which provided valuable insights into the dataset's structure and variability, attention was focused on the distribution of the target variable prime_estimee. Despite the overall data quality and completeness, initial exploration revealed an imbalance in pre- mium values, with some ranges particularly higher premiums being under-represented.

Such imbalances can adversely affect the performance and generalization of regression models by bi- asing predictions toward the majority target regions.

To address this, a data augmentation strategy was adopted using the SMOGN (Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise) algorithm.

SMOGN is specifically tailored for regression problems and synthetically generates new data points in the under-represented areas of the target variable while preserving the dataset's intrinsic structure and relationships. The table below presents descriptive statistics of key variables after applying SMOGN, illustrating the resulting dataset's enhanced balance.

Table 10: Descriptive Statistics of Selected Variables After SMOGN

Variable	Count	Mean	Median	Std	Min	Max
age	190	39.08	36.00	14.94	18.00	68.00
license_age	190	19.69	15.50	14.75	2.00	51.00
vehicle_age	190	5.81	6.00	3.45	0.00	17.00
power	190	6.62	6.00	1.61	4.00	10.00
passengers	190	5.00	5.00	0.00	5.00	5.00
Fuel Type_encoded	190	1.23	1.00	0.51	0.00	2.00
Gender_encoded	190	2.27	2.00	0.63	1.00	3.00

Estimated Premium	190	362.34	369.58	57.85	245.49	483.48

Following the detection of imbalance in the target distribution, the SMOGN data augmentation algorithm was applied to synthetically enhance under-represented target regions, resulting in a more balanced dataset.
The pseudo-code below summarizes the SMOGN implementation process.

Algorithm 4 Data Augmentation Using SMOGN for Regression

Require: Dataset D with features X and target variable y
Ensure: Augmented dataset D_{aug}
1: Load dataset D
2: Identify the target variable y to be balanced
3: Initialize SMOGN with dataset D and target y
4: Generate synthetic samples in under-represented regions of y using SMOGN
5: Combine synthetic samples with original data to form D_{aug}
6: Save or return the augmented dataset D_{aug}

3.6.5 Evaluation Data

After applying the SMOGN data augmentation technique to address the imbalance in the target variable distribution, the resulting augmented dataset exhibits improved balance and diversity, which is expected to enhance model training and generalization.
The subsequent evaluation phase involves assessing the predictive performance of the model trained on this enhanced dataset. To this end, a representative evaluation dataset processed and encoded in the same manner is used.
The table below presents a sample extract of this evaluation data, detailing key driver and vehicle features along with their corresponding estimated insurance premiums:

Table 11: Data Evaluation Table

Gender_encoded	age	vehicle_age	Power	combution_encoded	license_age	passengers	Estimated Premium
2	59	10	9	1	40	5	365.89
2	58	9	9	1	40	5	366.93
1	58	10	9	1	40	5	366.29
2	59	10	9	1	40	5	365.77
2	60	10	9	1	39	5	375.46
3	24	6	5	1	6	5	376.53
3	26	2	4	0	6	5	368.43
2	25	2	4	1	5	5	366.81
3	26	2	5	1	6	5	373.75
3	25	4	5	1	6	5	376.53
2	66	11	7	1	44	5	381.18
2	59	10	9	1	41	5	368.87
2	61	10	8	1	39	5	377.93
2	65	11	7	0	44	5	371.73
2	60	10	9	1	41	5	368.87

The table presents a sample extract of automobile insurance data, including several key variables that influence the estimated insurance premium.
Quantitative variables such as driver age, license tenure, vehicle age, engine power, and number of passengers are displayed with their corresponding values for each observation.
The categorical variables, including gender_encoded and fuel type_encoded, represent encoded cat- egories for driver gender and vehicle fuel type, respectively. In this extract, driver ages range from 24 to 66 years, with license tenures varying between 5 and 44 years.
Vehicle ages span from 2 to 11 years, while engine power ranges from 4 to 9 fiscal horsepower. All vehicles in the sample accommodate 5 passengers, reflecting a standard configuration.
The gender_encoded variable takes values between 1 and 3, indicating different gender categories, while fuel type_encoded uses numeric codes (1 or 2) to distinguish fuel types, with 2 likely represent- ing gasoline and 1 diesel. The estimated insurance premiums (prime_estimee) vary approximately

between 365.77 and 381.18 monetary units, reflecting the combined effect of the listed characteristics on the insurance cost. This extract highlights the comprehensive and processed nature of the dataset
used to model insurance premiums, integrating personal factors (age, gender, license tenure), technical vehicle attributes (age, power, fuel type), and environmental considerations. To gain deeper insights
into the interrelationships among these variables and their influence on the estimated premiums, a correlation analysis was performed.
Pearson's correlation coefficient was used as a statistical measure to quantify the strength and direc- tion of linear associations between the independent variables and the target variable prime_estimee. The following correlation matrix summarizes these relationships, providing an initial assessment of variable coherence and potential multicollinearity issues.

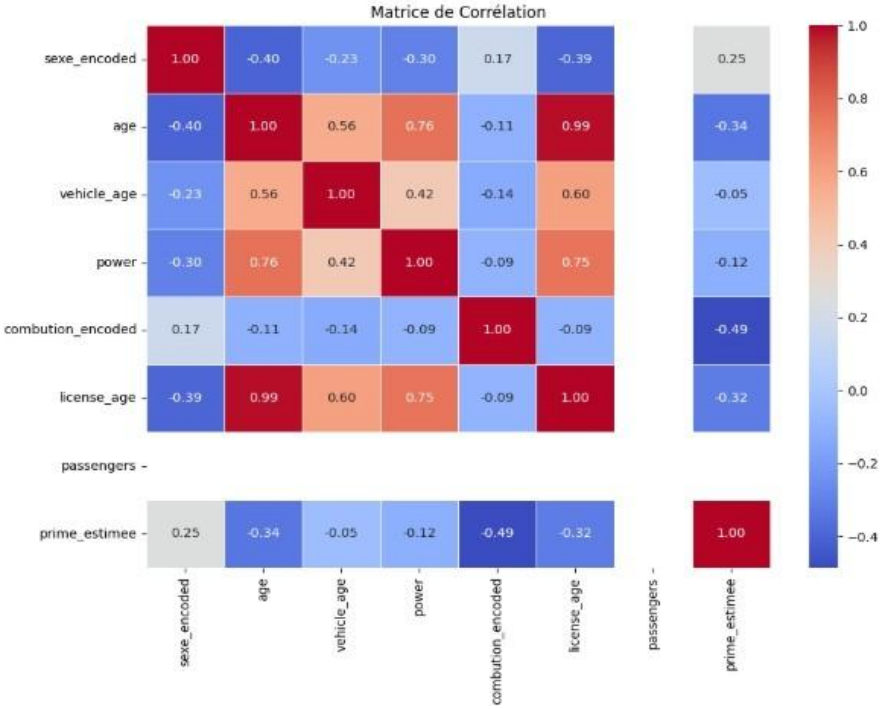

Figure 4: Correlation Matrix

To complement this quantitative analysis, additional visual examination of the pricing algorithm was performed using a real dataset, which included insured characteristics such as driver age, license se- niority, vehicle age, tax horse power, and number of seats. The resulting graphs provided valuable
insights into the behavior of each variable and its impact on the premium. It was observed that the premium decreases with increasing driver experience, reflecting the lower risk associated with more experienced drivers.
Conversely, the premium increases with vehicle age, number of seats, and tax horse power.
This can be explained by several factors: older vehicles are generally more prone to breakdowns, ve- hicles with more seats may be used commercially or carry higher liability, and higher-powered cars tend to have greater repair costs and are linked to increased risk exposure or claim frequency.
To complement the statistical and correlation-based insights, a series of graphical analyses was con-
ducted in order to visualize how each individual variable affects the estimated insurance premium. These visual tools provide an intuitive understanding of the underlying mechanisms driving the pricing algorithm and reinforce the reliability of the model when applied to real-world data. Following the
correlation analysis and initial descriptive insights, the next step involves standardizing the dataset to ensure that all numerical variables are on a comparable scale.
This normalization process is essential to enhance the performance and stability of subsequent statis- tical modeling and to allow a clearer interpretation of the numerical variable distributions.

3.6.6 Normalization and Numerical Variable Distribution

To ensure consistent scale across numerical variables and improve model performance, normalization was applied to the dataset.
This step is especially important for algorithms sensitive to variable magnitudes, such as distance- based or gradient-based models. In this study, Min-Max normalization was chosen, as it scales all
numerical features into a common range [0, 1] without distorting differences in value distributions. The normalization formula applied is as follows:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

The variables subjected to normalization include: age, license_age, vehicle_age, power, and prime_estimee. Categorical variables encoded numerically (gender_encoded, fuel type_encoded) were excluded from this normalization step to preserve their categorical meaning. Table 12 presents a sample of the normalized dataset.

Table 12: Data Evaluation Table (Normalized Variables)

Gender_encoded	age	vehicle_age	power	Fuel Type_encoded	license_age	passengers	Estimated Premium
0.50	0.82	0.59	0.83	0.78	0.00	0.51	0.50
0.50	0.80	0.53	0.83	0.78	0.00	0.51	0.50
0.00	0.80	0.59	0.83	0.78	0.00	0.51	0.50
0.50	0.82	0.59	0.83	0.78	0.00	0.51	0.50
0.50	0.84	0.59	0.83	0.76	0.00	0.55	0.50
1.00	0.12	0.35	0.17	0.08	0.00	0.55	0.50
1.00	0.16	0.12	0.00	0.08	0.00	0.52	0.00
0.50	0.14	0.12	0.00	0.06	0.00	0.51	0.50
1.00	0.16	0.12	0.17	0.08	0.00	0.54	0.50
1.00	0.14	0.24	0.17	0.08	0.00	0.55	0.50
0.50	0.96	0.65	0.50	0.86	0.00	0.57	0.50
0.50	0.82	0.59	0.83	0.80	0.00	0.52	0.50
0.50	0.86	0.59	0.67	0.76	0.00	0.56	0.50
0.50	0.94	0.65	0.50	0.86	0.00	0.53	0.00
0.50	0.84	0.59	0.83	0.80	0.00	0.52	0.50

The table above presents a sample of the normalized variables used in this study, where each quantitative variable has been transformed using Min-Max normalization to scale values between 0 and 1.
This preprocessing step is essential to standardize the range of features before applying statistical models or machine learning algorithms, preventing variables with larger scales from dominating those with smaller scales.

Table 13: Description of Normalized and Encoded Variables Used in the ANN Model

Variable	Description
Gender_encoded	Numerically encoded gender (0.00, 0.50, 1.00) to integrate into scaled models.
Age	Age of the insured person, normalized to represent relative youth or seniority.
Vehicle_age	Age of the vehicle, normalized to assess its impact on risk and premium.
Power	Fiscal horsepower of the vehicle, normalized to reduce scale-related bias.
Fuel_type_encoded	Encoded engine type (e.g., gasoline, diesel, electric), normalized for model input.
License_age	Driver's license tenure, normalized as a proxy for experience.
Passengers	Number of passengers, scaled relative to the dataset to observe potential risk impact.
Estimated Premium	Insurance premium estimate, normalized for consistent quantitative modeling.

This Min-Max normalization ensures that all variables contribute proportionally during subsequent analyses such as predictive modeling, regression, or classification algorithms.
This step is crucial to avoid bias caused by differing variable scales and to achieve more stable and performant models.
The illustrative table displays diverse profiles within the data, covering a broad range of values for each feature, thus ensuring good representativeness of the sample. This normalization process ensures that each input variable contributes proportionately to the learning process and avoids dominance effects caused by differing variable scales.

Algorithm 5 Min-Max Normalization of an Excel File

Require: Excel file path $path_{in}$, list of columns to normalize C, output CSV path $path_{out}$
Ensure: Normalized dataset saved as CSV file at $path_{out}$
1: Load Excel file from $path_{in}$
2: **if** loading fails **then**
3: Print error message and **stop**
4: **else**
5: Print ”File loaded successfully”
6: **end if**
7: Extract list of columns in dataset as Cols
8: Find missing columns: MissingCols $\rightarrow \{c \in C \mid c \notin Cols\}$
9: **if** MissingCols is not empty **then**
10: Print warning with MissingCols
11: **stop**
12: **end if**
13: Select columns C from dataset into D
14: Initialize Min-Max scaler
15: Apply scaler to D, producing normalized data D_{norm}
16: Print first 10 rows of D_{norm}
17: Save D_{norm} as CSV file to $path_{out}$
18: **if** saving fails **then**
19: Print error message
20: **else**
21: Print confirmation message
22: **end if**

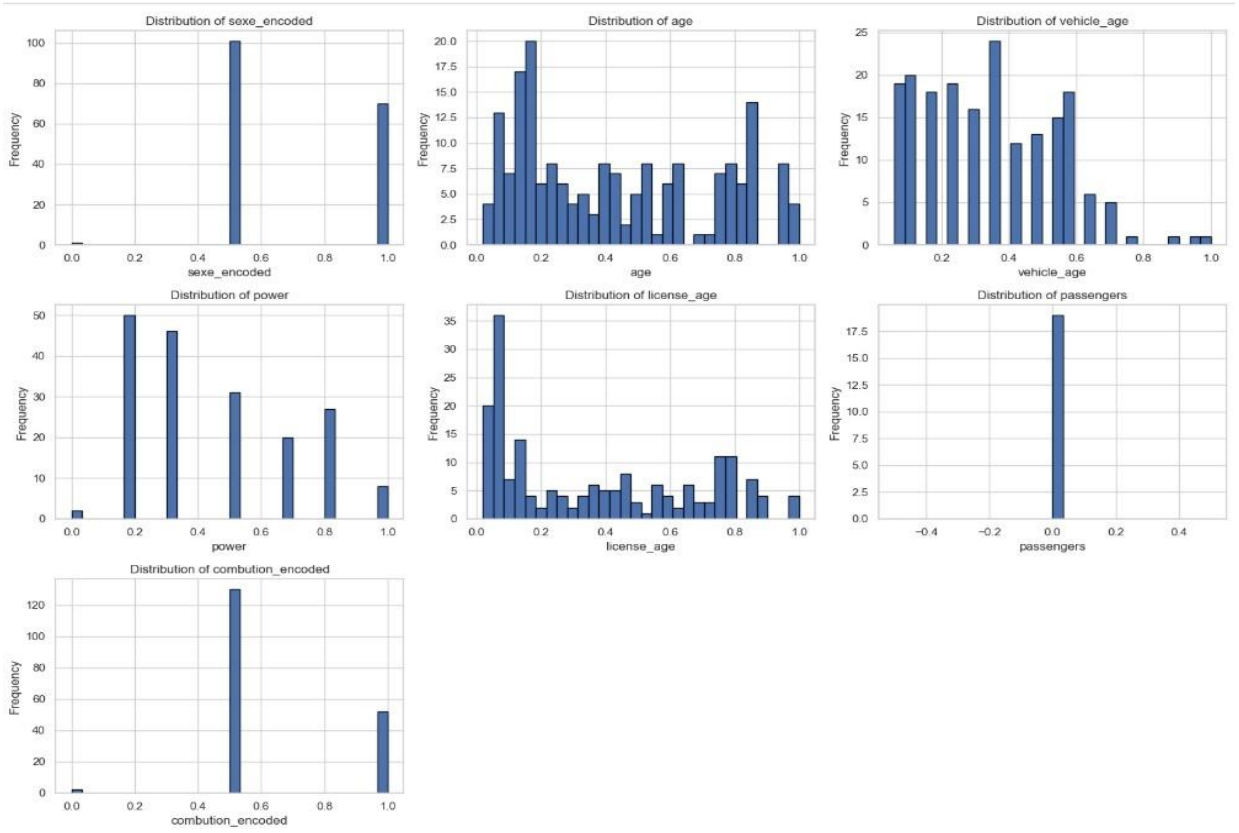

Figure 5: Graphical Analysis of Variable Effects on Insurance Premium

The following histograms illustrate the distribution of key variables within the dataset used for motor insurance pricing. These interpretations highlight important patterns and potential issues for modeling:

- **Distribution of gender_encoded:** This binary variable shows two clear peaks around 0 and 1 and 2, corresponding to the two categories of gender (Company=1, male= 2, female=3). A slight imbalance suggests an overrepresentation of one group, which could introduce bias if not accounted for.
- **Distribution of age:** The age variable spans the entire normalized range [0, 1], with several dispersed peaks.

This indicates a heterogeneous insured population in terms of age, which contributes positively to the representativeness of the dataset.

- **Distribution of vehicle_age:** A clear concentration in the lower part of the scale (0.1–0.4) suggests that most insured vehicles are relatively new.

A few older vehicles appear as outliers, which may be important to consider since vehicle age is correlated with claim risk.

- **Distribution of power:** The power variable shows distinct peaks, implying grouped categories (low, medium, high engine power).

Such clustering often reflects market segmentation and must be considered when pricing premi- ums based on vehicle specifications.

- **Distribution of license_age:** This variable is heavily skewed toward lower values, indicating a large proportion of drivers with recently obtained licenses. Since inexperienced drivers typi- cally present higher risk, this distribution is critical for risk assessment.
- **Distribution of passengers:** The near-constant value observed for passengers (with a sharp peak around 0) indicates a lack of variability.

This homogeneity in vehicle usage suggests that the variable may not provide substantial pre- dictive value in this dataset.

- **Distribution of fuel type_encoded:** Like gender_encoded, this binary variable shows a strong imbalance, indicating that most vehicles have the same fuel type (predominantly gasoline or diesel).

This lack of diversity may limit the variable's predictive power but remains relevant to include if it significantly affects risk.

To analyze the relationship between these factors and the target variable prime_estimee, scatter plots were drawn for each explanatory variable to observe trends and dispersion.

These plots allow for the detection of linear or nonlinear relationships, threshold effects, as well as the variability of the premium across different levels of the variables.

To explore the potential impact of each explanatory variable on the estimated insurance premium, scatter plots were generated for both numerical and encoded categorical variables.

These plots allow for a visual evaluation of possible trends, nonlinearities, or clusters that may inform the selection and transformation of features in later modeling stages.

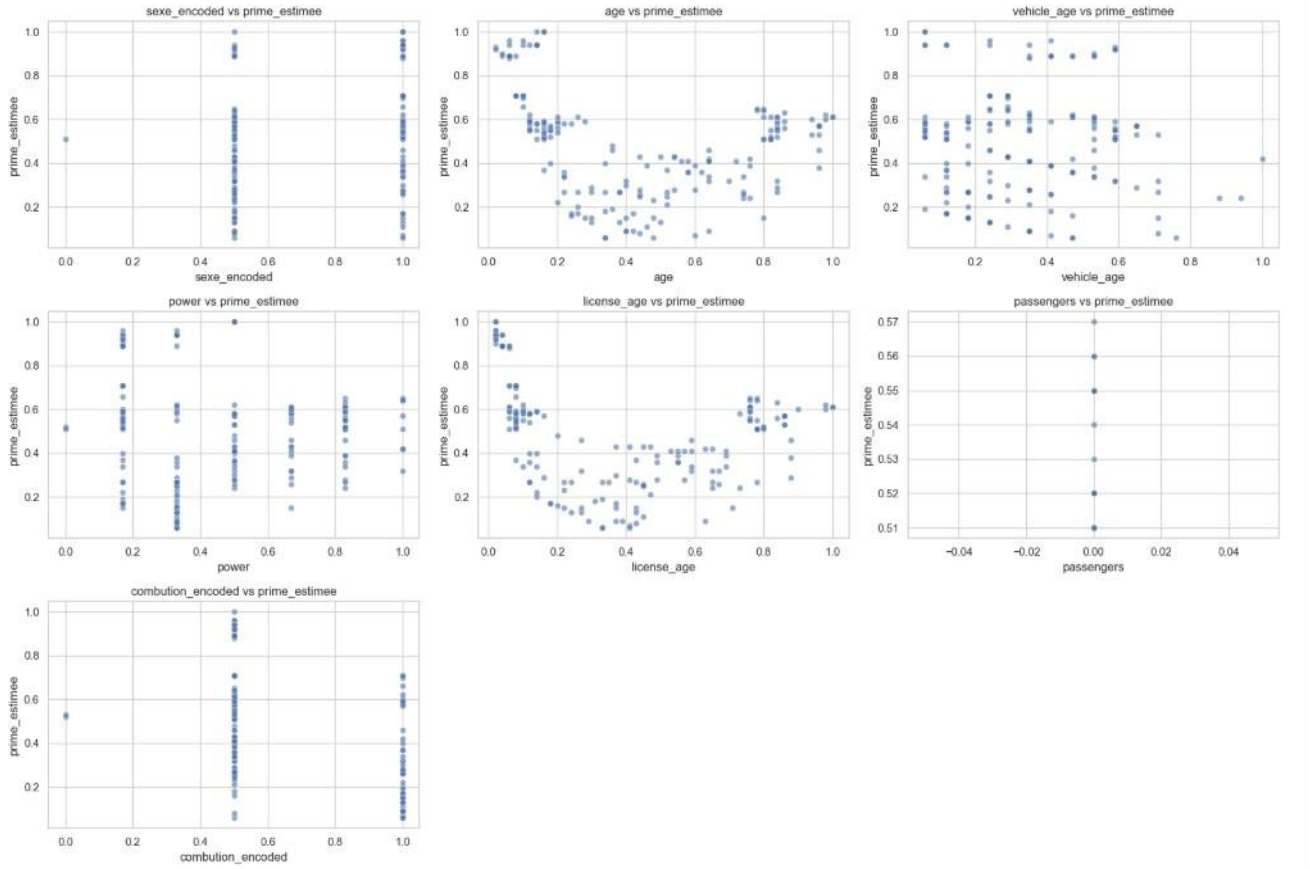

Figure 6: Scatter Plots of Predictor Variables Against Estimated Insurance Premium

(1) age: Generally shows a negative association with the estimated premium.
As the age of the insured increases, the premium tends to decrease, likely due to greater maturity and lower risk.
However, beyond a certain threshold (around age 60), the risk may increase again due to age-related decline in driving ability.
(2) license_age: Displays a strong negative correlation.
Less experienced drivers (with fewer years of driving) tend to be charged higher premiums due to higher perceived risk.
(3) power: Shows a dispersed pattern with no clear linear trend.
Nonetheless, higher engine power may be linked to riskier driving behaviors, potentially affecting premiums.
(4) Gender_encoded: Does not present a meaningful difference across categories. Premium distributions are similar for both genders.
(5) Fuel Type_encoded: Represents fuel type and does not exhibit a significant trend, suggesting limited impact on premium estimation.
(6) passengers: Appears constant across the dataset, offering no variability. This variable can be excluded from further modeling due to its lack of predictive value.
These graphical and methodological analyses demonstrate an integration between deep mathematical understanding and the behavior of real-world data, confirming the model's ability to simulate a reliable pricing logic that can be trusted in practical applications.
Incorporating these analyses within the premium calculation mechanism not only improves prediction accuracy but also enhances transparency and fairness in insurance product pricing.

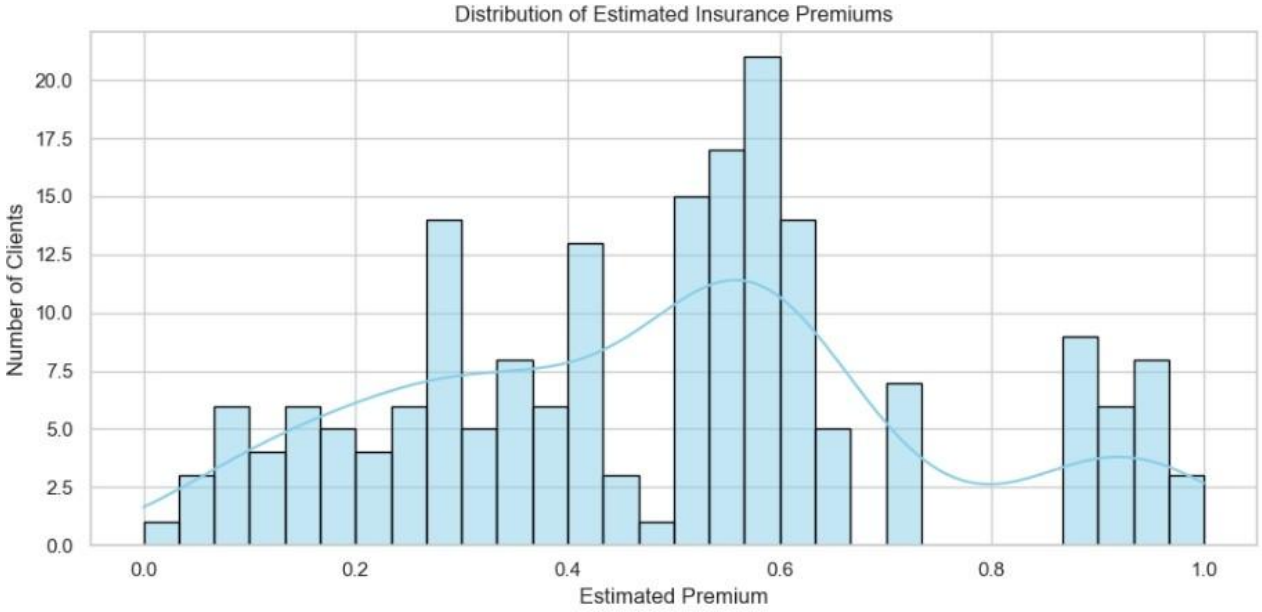

Figure 7: Annotated graph explaining the insurance premium

The pricing algorithm was applied to a representative real dataset containing information about insured individuals, including key variables such as the driver's age, years of license possession, vehicle age, fiscal power, and number of seats.
The application results yielded several indicators which were visually analyzed to understand the relationship between these variables and the insurance premium level.
The graphs shown illustrate the individual impact of the driver's age, license seniority, and vehicle age on the calculated premium.
It is evident from these graphs that the premium tends to decrease as driving experience increases, reflecting the relatively lower risk associated with more experienced drivers. Conversely, the premium increases with vehicle age, number of seats, and fiscal power, indicating a higher potential risk exposure or likelihood of claims.
These graphical analyses provide a visually supported understanding of the mathematical rationale underlying the premium adjustment factors used in the implemented algorithm.

3.7 Rationale for the Chosen Methodology

The quantitative mathematical methodology was adopted in this study due to its high efficiency in representing complex relationships among various factors influencing insurance pricing.
This methodology is distinguished by its ability to produce accurate and objective results based on real data, avoiding reliance on heuristic estimates or general assumptions.
Actuarial and probabilistic models are fundamental in insurance science, allowing for risk analysis and prediction of potential losses based on historical data, thus serving as an ideal tool for determining premiums reflecting the actual risk level of each insured.
Regarding optimization techniques, genetic algorithms were chosen for their effectiveness in handling nonlinear and multi-objective

optimization problems. They are particularly suitable in insurance pric- ing contexts requiring a precise balance between maximizing profits and minimizing potential risks.

These algorithms demonstrate notable efficiency in improving reinsurance strategies, enhancing their validity in auto insurance.

Python was also utilized for data analysis due to its advanced capabilities in managing large databases and its specialized libraries for numerical analysis and statistical modeling.

Python's flexibility in developing algorithms and testing multiple scenarios further improves the ac- curacy and realism of the models.

Overall, adopting this methodology aims to achieve integrated objectives, including: establishing fair pricing that reflects the individual characteristics and true risk level of each insured, improving insur- ance companies' technical performance by narrowing the gap between collected premiums and paid claims, and enhancing analytical models' ability to predict risks and adapt to future market fluctua- tions.

3.8 Pricing Formula Used in the Premium Calculation Algorithm

The adopted pricing formula is based on a mathematical model that accounts for the individual char- acteristics of each insured by integrating a set of risk factors into a flexible equation with customized weights for each variable.

The insurance premium for individual i is calculated according to the following equation:

$$\text{Premium}_i = x_{\text{base}} \times a_1 \left(F_{\text{license}} + a_2 \cdot F_{\text{vehicle}} + a_3 \cdot F_{\text{power}} + a_4 \cdot F_{\text{passengers}} + a_5 \cdot F_{\text{age}} + a_6 \cdot F_{\text{fuel}} + a_7 \cdot F_{\text{gender}} \right)$$

where:

x_{base}: unified base premium,

a_j: weight assigned to each risk factor, determined through statistical or actuarial analysis,

$F_{license}$: license seniority factor, $F_{vehicle}$: vehicle age factor, F_{power}: fiscal power factor,

$F_{passengers}$: number of passengers factor,

F_{age}: driver age factor, F_{fuel}: fuel type factor, F_{gender}: gender factor.

Each factor is computed as follows:

$$F_{\text{license}} = \max\left(0.85,\quad 1.6 - 0.06 \times A_{\text{license}}\right),$$
$$F_{\text{vehicle}} = \max\left(0.7,\quad 1.2 - 0.03 \times A_{\text{vehicle}}\right),$$
$$F_{\text{power}} = \min\left(1.8,\quad 1.0 + 0.07 \times P\right),$$
$$F_{\text{passengers}} = 1.0 + 0.03 \times \max\left(0,\quad N - 1\right),$$
$$F_{\text{age}} = \min\left(1.5,\quad 1.0 + 0.02 \times \max(0, 30 - A)\right),$$
$$F_{\text{fuel}} = \begin{cases} 1.0 & \text{if gasoline (2)} \\ 0.95 & \text{if diesel (1)} \end{cases},$$
$$F_{\text{gender}} = \begin{cases} 1.0 & \text{for Company (1)} \\ 1.1 & \text{for male (2)} \\ 1.05 & \text{for female (3)} \end{cases}$$

$A_{license}$: years of license possession,

$A_{vehicle}$: vehicle age in years,

P: vehicle fiscal power,

N: number of passengers in the vehicle,

A: age of the driver,

Fuel (1 = Diesel, 2 = Gasoline),

Gender (1 = Company, 2 = Male, 3 = Female).

This estimation formula reflects the model's ability to adapt to individual traits of each insured, contributing to a fairer and more accurate pricing based on the actual risk level associated with each insurance case.

The following algorithm translates the pricing formula from Section 3.8 into a practical computation procedure.

It receives as inputs the driver's and vehicle's characteristics, determines the corresponding risk factors, and combines them to produce the optimal insurance premium.

By following a clear, step-by-step process, it guarantees consistency, transparency, and adaptability to different profiles.

Each factor such as driver age, license seniority, vehicle age, and usage type is computed according to predefined actuarial rules. The final premium is obtained by multiplying the base civil liability cost by the combined effect of all risk factors, ensuring accurate and fair pricing for each policyholder.

3.9 The Algorithm: Calculating the Optimal Car Insurance Premium

Algorithm 6 Estimated Insurance Premium Calculation

Require: Driver's age, License age, Vehicle age, Horsepower, Number of passengers, Usage type
Ensure: Return estimated insurance premium
civil_liability $\rightarrow$ 300 ▷ Base cost for civil liability
if age $<$ 25 **then**
age_factor $\rightarrow 1.6 - 0.02 \times (25 - \text{age})$
else if age $\leq$ 60 **then**
age_factor $\rightarrow 1.0 + 0.01 \times (30 - |\text{age} - 30|)$
else
age_factor $\rightarrow 1.3 + 0.015 \times (\text{age} - 60)$
end if
license_factor $\rightarrow \max(0.85,\ 1.6 - 0.06 \times \text{license\_age})$
vehicle_factor $\rightarrow 1.0 + 0.04 \times \min(\text{vehicle\_age},\ 15)$
power_factor $\rightarrow 1.0 + 0.015 \times$ power
passenger_factor $\rightarrow 1.0 + 0.03 \times \max(0,\ \text{passengers} - 1)$
if usage_type = "personal" **then**
usage_factor $\rightarrow$ 1.0
else if usage_type = "professional" **then**
usage_factor $\rightarrow$ 1.4
else
usage_factor $\rightarrow$ 1.15
end if
multiplier $\rightarrow$ age_factor $\times$ license_factor $\times$ vehicle_factor $\times$ power_factor $\times$ passenger_factor $\times$ usage_factor
total $\rightarrow$ civil_liability $\times$ multiplier
return total

Algorithm 6 represents the core computational mechanism for implementing the optimal reinsurance pricing model developed in this study. It estimates the insurance premium by systematically adjusting a base civil liability cost according to several key risk factors related to the driver, the vehicle, and usage characteristics.

The algorithm begins with a fixed baseline premium for civil liability coverage.

This base is then modified through multiplicative coefficients reflecting individual risk determinants such as the driver's age, license tenure, vehicle age, fiscal horsepower, number of passengers, and usage type (personal or professional).

This multiplicative approach provides a flexible and dynamic framework for individualized risk as- sessment, enabling the calculation of tailored premiums that correspond to the unique profile of each insured entity.

Its simplicity and adaptability make it suitable for practical deployment in insurance pricing systems. Furthermore, by generating precise quantitative estimates, the algorithm supports integration into more advanced optimization procedures such as augmented Lagrangian methods or genetic algorithms facilitating enhanced reinsurance strategy design.

Additionally, this computational tool aids sensitivity analyses, allowing insurers to understand how parameter variations affect premium calculations and overall risk management decisions.

4 RESULTS AND DISCUSSION

The developed feedforward artificial neural network (ANN) was trained to estimate insurance premi- ums based on seven input features, using a normalized dataset enhanced via the SMOGN technique. The model architecture consisted of a single hidden layer containing four neurons with ReLU activa- tion, and an output layer with linear activation for regression.

Training was performed over 100 epochs with early stopping applied, halting at epoch 67 to prevent overfitting.

The model used the Adam optimizer and Mean Squared Error (MSE) as the loss function. A summary of model characteristics and training outcomes is presented in Table 14.

Table 14: Summary of Model Results

Metric	Value	Description
Model type	Feedforward ANN	Dense neural network with 1 hidden layer
Input features	7	Gender_encoded, age, vehicle_age, power, license_age, passengers, Fuel_type_encoded
Hidden layer neurons	4	Activation: ReLU
Output layer neurons	1	Activation: Linear
Loss function	MSE	Mean Squared Error
Optimizer	Adam	Adaptive optimizer
Epochs trained	67	Early stopping applied
Batch size	16	–
MAE (test set, normalized)	0.625	Mean Absolute Error on normalized target

The model described above is a relatively simple feedforward artificial neural network (ANN), consisting of a single hidden layer with four neurons and ReLU activation.

This architecture was deliberately chosen to maintain interpretability and avoid the complexity of deeper models, which could lead to overfitting given the limited dataset size.

The output layer employs a linear activation function, in line with the regression objective of predict- ing continuous insurance premiums.

The preprocessing pipeline involved normalization and categorical encoding essential steps in ANN modeling to ensure proper convergence and eliminate scale-related biases.

A total of seven input features were selected based on their theoretical relevance to insurance risk (driver age, vehicle power, driving experience), enhancing the model's explanatory capacity.

Training was performed using the Adam optimizer, widely recognized in deep learning for its adaptive learning rate and robustness to sparse gradients. To mitigate overfitting, early stopping was imple- mented based on the validation loss.

The model converged at epoch 67, indicating an optimal trade-off between underfitting and overfit- ting. With a normalized Mean Absolute Error (MAE) of 0.625, the model demonstrates a reasonable ability to generalize in scaled terms; however, further assessment on the original scale of the target variable is necessary for a complete performance evaluation.

Table 15: Performance Metrics of the ANN Model

Metric	Value
Mean Absolute Error (MAE)	26.94
Mean Squared Error (MSE)	1225.46
Root Mean Squared Error (RMSE)	35.01
Coefficient of Determination (R^2)	0.65

Table 15 illustrates the performance metrics derived from *denormalized* predictions, providing a concrete evaluation of the model's accuracy in estimating insurance premiums.

The Mean Absolute Error (MAE) of **26.94** indicates that, on average, the predicted premiums deviate from the actual premiums by approximately 27 monetary units.

Given the inherent complexity and nonlinearity of insurance risk factors, this level of average deviation reflects the model's reasonable precision in pricing, effectively balancing bias and variance.

The Mean Squared Error (MSE) of **1225.46** and the associated Root Mean Squared Error (RMSE = **35.01**) highlight the dispersion of errors around the mean prediction.

The moderate gap between MAE and RMSE suggests that the model is not unduly influenced by extreme outliers or large errors; rather, it maintains consistent predictive reliability across diverse cases, which is critical for actuarial applications sensitive to significant losses.

The coefficient of determination (R^2 = **0.65**) indicates that 65% of the variance in observed premium values is explained by the model. This level of explanatory power is noteworthy, considering the heterogeneity and multifactorial nature of the insurance datasets.

Table 16: Sample of Actual vs Predicted Premiums

Real Premium	Predicted Premium
258.72	309.05
306.98	324.18
354.66	346.69

456.73	406.96
349.00	364.20

Table 16 demonstrates the ANN model's capability to approximate the underlying pricing mech- anism of automobile insurance premiums by leveraging structured input features.

Future research could explore enhancements such as increasing the number of hidden layers, fine- tuning hyperparameters, or employing ensemble methods to further improve predictive performance. In this study, a Generalized Linear Model (GLM) with a Poisson distribution was also developed to estimate insurance premiums based on explanatory variables related to the insured individual and the vehicle.

The model incorporates factors such as driver age, years of license possession, vehicle age, engine power, number of seats, and fuel type. Statistical analyses were conducted using the statsmodels library in Python, and the model's accuracy was assessed through goodness-of-fit statistics and esti- mation quality metrics.

To gain further insights into the relationships between individual risk factors and insurance pre- miums, the study also applies a Generalized Linear Model (GLM) with a Poisson distribution.

This approach offers a more interpretable, statistically grounded framework, allowing us to examine the significance and direction of effects for each explanatory variable on the premium estimation.

The following section presents the GLM results, including coefficient estimates, statistical significance, and model performance metrics, complemented by visual comparisons between predicted and actual premiums.

Table 17: GLM Results Using Poisson Distribution

Generalized Linear Model Regression Results

Dep. Variable:	prime_estimee	No. Observations:	190
Model:	GLM	Df Residuals:	183
Model Family:	Poisson	Df Model:	6
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1175.4
Date:	Tue, 05 Aug 2025	Deviance:	884.34
Time:	23:53:42	Pearson chi2:	875.
No. Iterations:	4	Pseudo R-squ. (CS):	0.9891
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.2341	0.002	137.066	0.000	0.231	0.237
sexe_encoded	0.0550	0.007	8.172	0.000	0.042	0.068
age	-0.0118	0.002	-6.976	0.000	-0.015	-0.008
license_age	0.0055	0.002	3.093	0.002	0.002	0.009
vehicle_age	0.0051	0.001	3.503	0.000	0.002	0.008
power	0.0315	0.004	8.646	0.000	0.024	0.039
passengers	1.1704	0.009	137.066	0.000	1.154	1.187
combution_encoded	-0.1712	0.008	-22.238	0.000	-0.186	-0.156

The results of the Generalized Linear Model (GLM) using a Poisson distribution reveal statistically significant relationships between several predictor variables and the estimated insurance premium:

- The variable **driver's sex** (Gender_encoded) has a **positive and statistically significant** effect on the premium ($P < 0.001$).

This indicates that premiums vary depending on the driver's gender, with the encoded category associated with a higher expected premium.

- The **driver's age** is **negatively correlated** with the estimated premium and highly significant ($P < 0.001$).

This suggests that older drivers tend to pay lower premiums, likely due to their reduced risk profile and more cautious driving behavior.

- The variable **license age** (seniority) has a **positive and statistically significant** effect on premiums ($P = 0.002$).

Contrary to expectations, this indicates that drivers with more years of holding a license may pay slightly higher premiums, potentially due to confounding effects or multicollinearity with age.

- The **vehicle age** also shows a **positive and significant** relationship with premiums ($P < 0.001$), which contrasts with the traditional expectation that older cars would result in lower premiums.

This may reflect repair costs or usage patterns associated with vehicle age in this dataset.

- **Engine power** exhibits a **positive and highly significant** influence on the premium ($P < 0.001$).

More powerful vehicles are generally considered riskier, leading to increased premiums due to higher potential accident severity.

- The number of **passengers** is associated with a **strong positive and significant** coefficient ($P < 0.001$). Vehicles that carry more passengers might pose higher liability exposure, justifying higher insurance costs.
- The **fuel type** , encoded as a dummy variable, shows a **negative and significant** effect on premiums ($P < 0.001$).

This indicates that certain fuel types (diesel or gasoline, depending on encoding) are associated with lower premium estimates.

Table 18: Performance Metrics of the Poisson GLM Model

Metric	Value
Mean Absolute Error (MAE)	33.65
Mean Squared Error (MSE)	1689.92
Root Mean Squared Error (RMSE)	41.11
Coefficient of Determination (R^2)	0.49

Performance metrics for the model are as follows: Mean Absolute Error (MAE) of 33.65, Mean Squared Error (MSE) of 1689.92, Root Mean Squared Error (RMSE) of 41.11, and coefficient of de- termination $R^2 = 0.49$, indicating a moderate predictive accuracy.

This suggests that the model captures important patterns in the data but still has room for improve- ment in predicting insurance premiums more precisely.

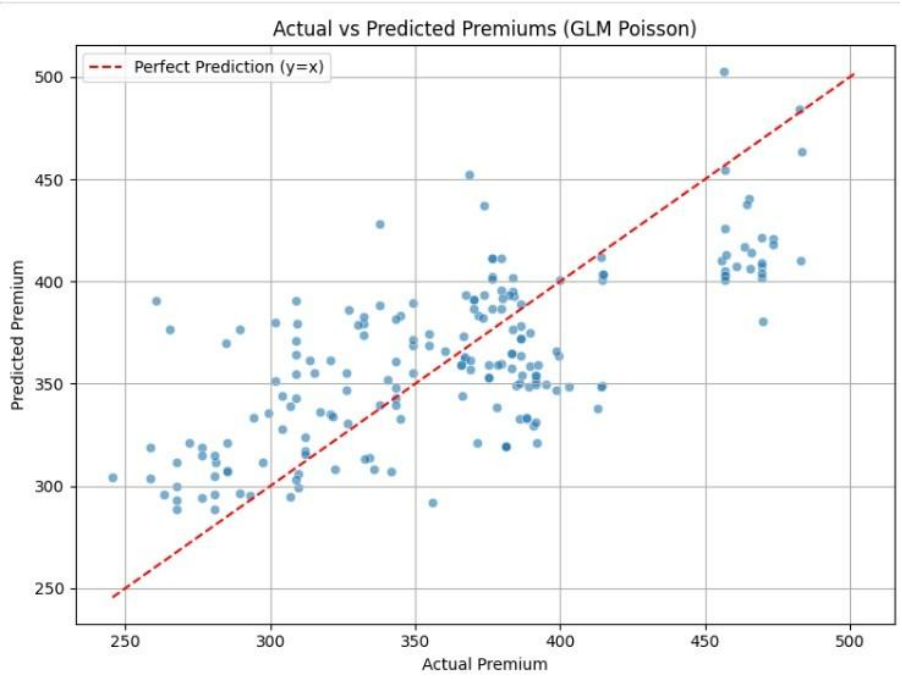

Figure 8: Comparison of Actual and Estimated Premiums Using Poisson Model

Most data points cluster near the identity line, indicating the model's ability to predict premiums with acceptable accuracy. However, some variance appears in higher premium values, where the model occasionally underesti- mates the actual premium. This behavior is considered acceptable in statistical models, particularly with a relatively high pseudo R-squared value (here, 0.49). Therefore, this plot supports the model's quality and demonstrates the closeness of estimates to real data, enhancing insurers' confidence in using the model for fair and precise pricing.

While the Poisson GLM provided valuable insights and a reasonable fit, it is important to address potential overdispersion present in the insurance premium data. To account for this, a Generalized Linear Model using a Negative Binomial distribution was also estimated.

The following section presents the results of this alternative model, which aims to better capture data variability and improve estimation robustness under conditions where the Poisson assumptions may be violated.

Table 19: GLM Results Using Negative Binomial Distribution

Generalized Linear Model Regression Results

Dep. Variable:	prime_estimee	No. Observations:	190
Model:	GLM	Df Residuals:	183
Model Family:	NegativeBinomial	Df Model:	6
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1308.7
Date:	Wed, 06 Aug 2025	Deviance:	2.4588
Time:	00:12:03	Pearson chi2:	2.40
No. Iterations:	4	Pseudo R-squ. (CS):	0.01245
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.2338	0.032	7.267	0.000	0.171	0.297
sexe_encoded	0.0601	0.127	0.472	0.637	-0.189	0.309
age	-0.0118	0.032	-0.371	0.710	-0.074	0.050
license_age	0.0058	0.033	0.176	0.860	-0.059	0.071
vehicle_age	0.0046	0.028	0.162	0.871	-0.050	0.060
power	0.0314	0.069	0.453	0.651	-0.105	0.167
passengers	1.1689	0.161	7.267	0.000	0.854	1.484
combution_encoded	-0.1774	0.146	-1.216	0.224	-0.463	0.109

Table 19 shows the results of the generalized linear regression model using a Negative Binomial distribution.
This model was chosen due to the presence of overdispersion in the premium values, which the tradi- tional Poisson model cannot efficiently handle.
The sample contained 190 observations, with 183 degrees of freedom remaining.
A log-link function was used to fit the dependent variable (premium), which only takes positive values. The log-likelihood value was −1308.7, reflecting a lower fit compared to the Poisson model.
The pseudo R-squared was also low at 0.01245, suggesting that the model explains only a small pro- portion of the total variance.
Analyzing the model coefficients, none of the input variables **driver's sex** (gender_encoded), **age**, **license seniority**, **vehicle age**, and **engine power** were statistically significant at the 5% level, as their P-values exceed 0.05.
This indicates that their individual effects are not substantial within this model. However, the number of **passengers** shows a strong positive and highly significant effect ($P < 0.001$), indicating a mean- ingful influence on premium estimation. The variable **fuel type** (encoded as fuel type_encoded) shows a negative but statistically insignificant effect ($P = 0.224$).

Table 20: Performance Metrics of the Negative Binomial GLM

Metric	Value
Mean Absolute Error (MAE)	33.79
Mean Squared Error (MSE)	1701.11
Root Mean Squared Error (RMSE)	41.24
Coefficient of Determination (R^2)	0.49

The model's performance metrics are summarized in Table 20, which reports a Mean Absolute Error (MAE) of 33.79, Mean Squared Error (MSE) of 1701.11, Root Mean Squared Error (RMSE) of 41.24, and a coefficient of determination (R^2) of 0.49, indicating moderate predictive accuracy.
To further assess the model's accuracy, a graphical comparison between actual and predicted premiums was generated (see Figure **??** below).

This visualization illustrates the closeness of the model's estimates to observed values, offering valuable insight into its effectiveness in simulating insurance premiums.

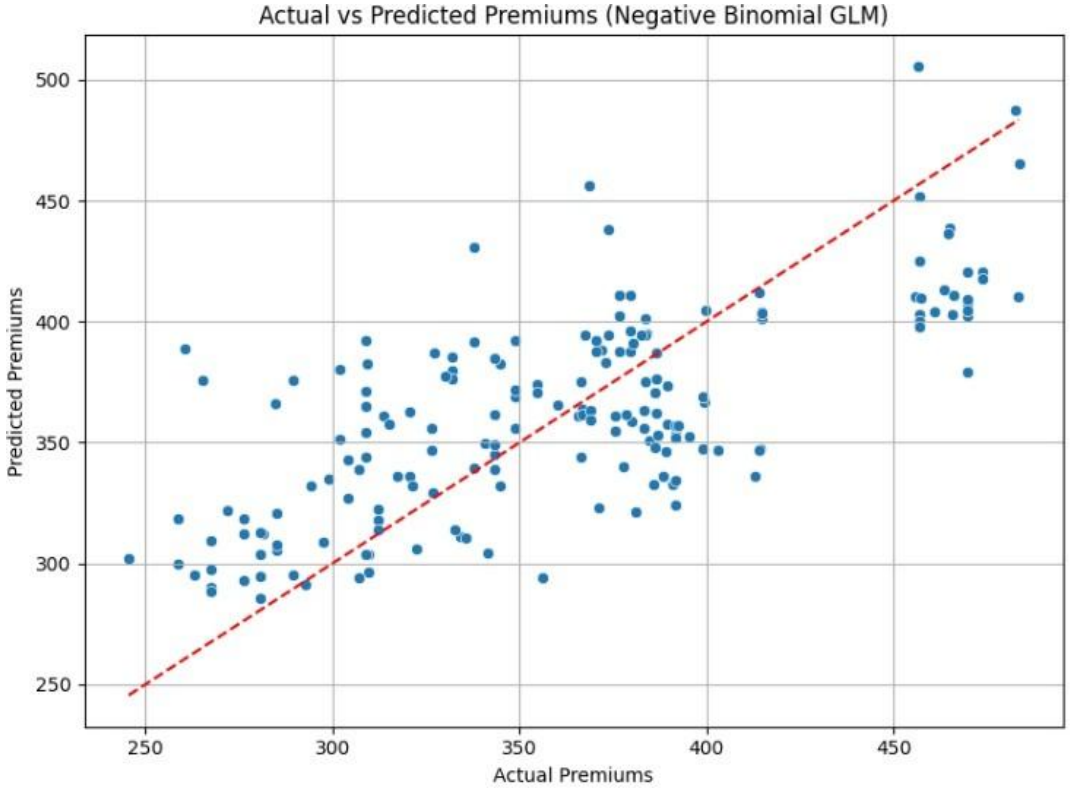

Figure 9: Graphical Comparison of Actual and Estimated Premiums Using the Generalized Linear Model with Negative Binomial Distribution

The graphical comparison between actual and estimated premiums shows that most data points lie close to the red diagonal line, which represents perfect equality between predicted and observed values.

This indicates that the model demonstrates a moderate to good predictive performance overall. However, a noticeable dispersion is observed in the upper range of premium values, where the model tends to **underestimate the actual premiums**.

This underestimation may be attributed to the absence of influential explanatory variables or the presence of nonlinear relationships not captured by the current model.

Despite the relatively low pseudo R-squared value, the plot suggests that the model is capable of capturing the general trend in the data.

Therefore, it can be considered a reasonable starting point for premium estimation.

Further improvements could be achieved by incorporating interaction effects, adding categorical pre- dictors, or adopting more flexible modeling techniques such as random forests or gradient boosting. Despite the relatively low pseudo R^2 values, the plots suggest that the models are capable of capturing the general trend in the data.

Therefore, these models can be considered reasonable starting points for premium estimation. Further improvements could be achieved by incorporating interaction effects, adding categorical pre-dictors, or adopting more flexible modeling techniques such as random forests or gradient boosting. While the graphical analysis provides valuable insights into each model's fit to the observed data, a more detailed comparison of their predictive performances is necessary to fully evaluate their effec- tiveness.

Table 21 presents a comparative overview of the performance metrics obtained from the different models applied to estimate insurance premiums.

The Artificial Neural Network (ANN) clearly outperforms the classical generalized linear models (Pois- son and Negative Binomial GLM), displaying lower error measures (MAE, MSE, RMSE) and a higher coefficient of determination ($R^2 = 0.65$). This superior performance indicates that the ANN more effectively captures the complex, nonlinear relationships inherent in the data, resulting in more accu- rate premium forecasts.

Nevertheless, the GLM approaches retain their value due to their interpretability, ease of implemen- tation, and relevance when linear relationships are assumed or when explanatory inference is required.

Table 21: Comparison of Model Performance Metrics

Model	MAE	MSE	RMSE	R^2
Poisson GLM	33.65	1689.92	41.11	0.49
Negative Binomial GLM	33.79	1701.11	41.24	0.49
Artificial Neural Network (ANN)	26.94	1225.46	35.01	0.65

While performance metrics such as MAE, RMSE, and R^2 provide insight into each model's ability to estimate average premiums, they do not capture the full spectrum of risk particularly in extreme loss scenarios. In the context of insurance, where rare but severe claims can significantly impact profitability and solvency, it becomes essential to complement traditional accuracy metrics with robust risk measures.

To this end, the next section introduces advanced risk indicators namely Value at Risk (VaR) and Tail Value at Risk (T-VaR) to evaluate the models' performance from a tail risk perspective.

These measures offer deeper insights into the potential financial exposure associated with each model's predictions, reinforcing their practical relevance for actuarial pricing and capital management.

In addition to evaluating predictive models using traditional metrics (MAE, MSE, RMSE, and R^2), this study incorporates advanced risk measures namely Value at Risk (VaR) and Tail Value at Risk (T-VaR) to assess the potential variability and extremity of estimated premiums.

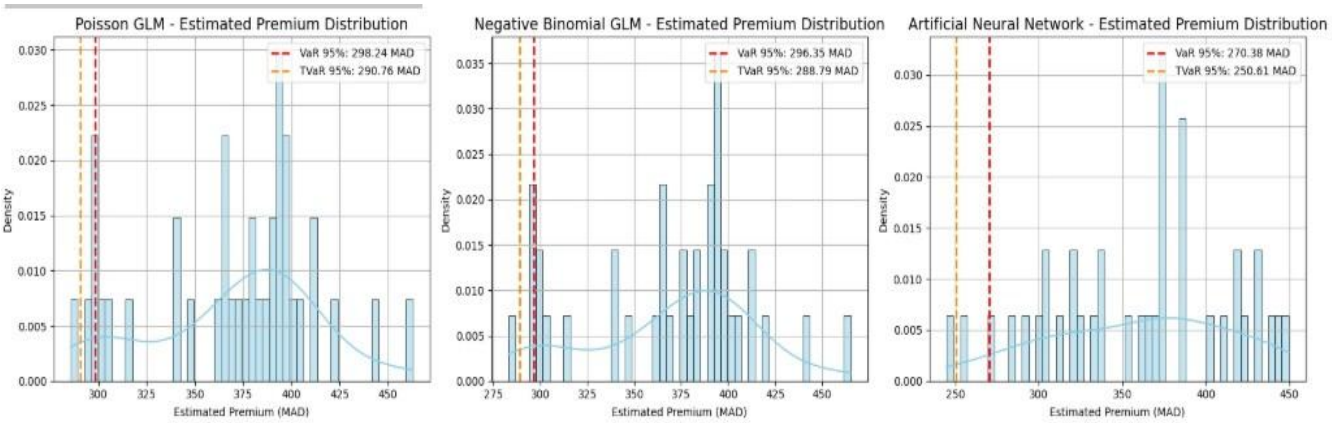

Figure 10: Distribution of Estimated Premiums Using VaR and T-VaR Risk Measures

Figure 10 illustrates the estimated premium distributions produced by three predictive models: the Poisson GLM, the Negative Binomial GLM, and the Artificial Neural Network (ANN).

Each subplot includes overlays of the Value at Risk (VaR) and Tail Value at Risk (T-VaR) at the 95% confidence level, marked by dashed vertical lines.

The **Poisson GLM** model presents a VaR of approximately 298.24 MAD and a T-VaR of 290.76 MAD.

Its distribution features a moderate peak around typical premium values but exhibits heavier tails, indicating a higher likelihood of extreme losses.

Similarly, the **Negative Binomial GLM** shows a comparable distribution pattern with slightly lower VaR (296.35 MAD) and T-VaR (288.79 MAD) values.

This suggests marginally better control over tail risk compared to the Poisson model.

In contrast, the **Artificial Neural Network (ANN)** distribution is broader and more symmetric, with significantly lower VaR (270.38 MAD) and T-VaR (250.61 MAD) values.

This reflects improved handling of extreme premium predictions and a substantial reduction in tail risk.

Table 22: Comparative Analysis of Model Performance and Risk Metrics

Model	VaR (95%)	T-VaR (95%)
Poisson GLM	298.24	290.76
Negative Binomial GLM	296.35	288.79
Artificial Neural Network (ANN)	**270.38**	**250.61**

As shown in Table 22, the Artificial Neural Network (ANN) consistently achieves the lowest VaR and T-VaR values among the models considered.

Specifically, the ANN reports a VaR of 270.38 MAD and a T-VaR of 250.61 MAD, which are notably lower than those of the Poisson and Negative Binomial GLMs. This improvement suggests that the ANN not only enhances the accuracy of central premium estimates but also better captures the tail risk linked to extreme loss events, thereby strengthening overall risk containment.

Conversely, the higher tail risk metrics observed for the Poisson and Negative Binomial models imply a greater expected severity of losses under adverse scenarios.

From a managerial standpoint, these results highlight the advantages of adopting ANN-based models for premium pricing, especially when controlling tail risk is paramount.

With growing regulatory demands on solvency and stress testing, integrating risk measures like T-VaR into actuarial assessments promotes pricing strategies that are more resilient, equitable, and financially robust.

Ultimately, this analysis advocates a shift from traditional, interpretable models towards flexible, risk- aware machine learning approaches.

Such models empower insurers to more effectively predict and manage financial exposures during ex- treme events, thereby enhancing solvency, ensuring fairness to policyholders, and supporting strategic robustness.

CONCLUSION

This research proposed a comprehensive and innovative framework for automobile insurance pricing in the Moroccan market, combining classical actuarial techniques with advanced artificial intelligence approaches. The methodology integrated Generalized Linear Models (GLMs), Artificial Neural Net- works (ANNs), to build a pricing system capable of delivering accurate, fair, and adaptive premiums tailored to the individual risk profile of each policyholder, while respecting technical and economic constraints.

From a methodological perspective, the Poisson and Negative Binomial GLMs effectively identified and quantified the influence of key rating factors, including driver's age, license seniority, vehicle power, and age of the vehicle. These models provided interpretability and statistical robustness, of- fering clear insights into the main determinants of premium variation.

The integration of ANNs significantly improved predictive accuracy by capturing nonlinear relation- ships and complex dependencies that classical models tend to overlook.

This enhancement was particularly evident in the reduction of extreme prediction errors, as demon- strated by lower Value at Risk (VaR) and Tail Value at Risk (T-VaR) values.

A major contribution of this work was the simulation of optimal reinsurance strategies using a hybrid genetic algorithm.

The results confirmed the effectiveness of a Stop-Loss contract in mitigating the impact of rare but severe losses.

By integrating quantitative risk indicators such as VaR and T-VaR into both pricing and reinsurance decisions, the study advances a practical approach to more resilient and solvency-oriented risk man- agement practices in the Moroccan insurance context.

Despite these achievements, the research has certain limitations. The dataset, while realistic and representative of the local market, was limited in size and lacked potentially relevant variables such as detailed claims history, telematics-based driving behavior, and environmental risk factors.

Furthermore, although machine learning models like ANNs enhance predictive capabilities, their com- plexity and opacity may hinder regulatory approval and adoption in operational insurance settings. Future research could extend this work by applying the methodology to other insurance domains such as health, property, or agricultural insurance to evaluate robustness across different risk profiles.

The exploration of advanced deep learning architectures, including recurrent and reinforcement learn- ing models, could further improve dynamic and adaptive pricing capabilities.

Moreover, conducting sensitivity analyses on incomplete, imbalanced, or noisy datasets would enhance model resilience.

Incorporating macroeconomic, demographic, and climate related variables could also enrich the anal- ysis by capturing systemic and long term risk factors.

Finally, the use of explainable AI techniques would improve the interpretability of neural network outputs, facilitating greater trust and acceptance among decision-makers.

In conclusion, this study demonstrates that the integration of actuarial science with artificial intelli- gence can significantly improve both the accuracy and the fairness of premium pricing in the Moroccan automobile insurance sector.

By combining statistical rigor, predictive modeling, and optimization strategies, the proposed ap- proach provides a solid foundation for developing data-driven, equitable, and strategically robust pricing systems.

Such innovations can strengthen insurers' solvency, enhance customer satisfaction, and foster sustain- able competitiveness in an increasingly regulated and complex market.

REFERENCES

1. Anisetti, M., Ardagna, C. A., Bena, N., & Foppiani, A. (2021). An assurance-based risk management framework for distributed systems. *2021 IEEE International Conference on Web Services (ICWS)*, 482–492.
2. Antonio, K., & Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *ASTIN Bulletin*.
3. Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin*, *3*(3), 199–207.
4. Bühlmann, H., & Gisler, A. (2005). *A course in credibility theory and its applications*. Springer.
5. Cai, J., & Tan, K. S. (2000). Conditional tail expectation for elliptical distributions. *Insurance: Math- ematics and Economics*, *27* (3), 345–356.
6. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-dm 1.0: Step-by-step data mining guide*. SPSS Inc.
7. Charpentier, A. (2020). *Data science for actuaries*. Springer.
8. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
9. de Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.
10. Denuit, M., et al. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. Wiley.
11. El Attar, M., et al. (2019). Application of neural networks to automobile insurance pricing. *Revue Marocaine de Recherche en Management et Marketing*.
12. Frees, E. W. (2010). *Regression modeling with actuarial and financial applications*. Cambridge Uni- versity Press.
13. Gerber, H. U. (1972). Games of economic survival with insurance applications. *ASTIN Bulletin*, *6*(2), 111–135.
14. Gong, Y., Li, Z., Milazzo, M., Moore, K., & Provencher, M. (2018). Credibility methods for individual life insurance. *Risks*, *6*(4), 144.
15. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.
16. Herzog, T. N. (1999). *Introduction to credibility theory*. Actex Publications.
17. Janssens, D., Wets, G., Brijs, T., Vanhoof, K., & Vanthienen, J. (2006). Integrating neural networks for car insurance fraud detection. *Expert Systems with Applications*, *29*(3), 653–666.
18. Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). *Modern actuarial risk theory: Using r* [Application directe de VaR et TVaR dans un contexte actuariel avec implémentation R]. Springer.
19. Klugman, S. A. (1992). Loss models and their estimation. *Insurance: Mathematics and Economics*, *11*(2), 105–121.
20. Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). *Loss models: From data to decisions* (4th).Wiley.
21. Liu, Y., & Shih, Y.-H. (2011). Integrating data mining and behavioral scoring for risk assessment of insurance customers. *Expert Systems with Applications*, *38*(6), 6551–6560.
22. Mahler, H. C., & Dean, C. G. (1999). Credibility (chapter 8) [Excellent chapitre appliqué pour les praticiens]. *Foundations of Casualty Actuarial Science, 4th Edition*.
23. McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd). Chapman; Hall/CRC.
24. Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models* (Vol. 174). Springer.
25. Rouyan, Y., & Amrani, S. (2020). Étude des déterminants de changement d'assureur automobile au maroc. *Revue Marocaine de Management*, *10*(1), 77–92.
26. Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Ware- housing*, *5*(4), 13–22.
27. Tan, K. S., & Weng, C. (2006). Value at risk and conditional tail expectation for actuarial risks: A comparative study. *North American Actuarial Journal*, *10*(4), 100–114.
28. Taylor, G. C., & Ashe, D. (1983). Credibility, exposure and risk measures in insurance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.
29. Wang, S., Young, V. R., & Panjer, H. H. (1998). Axiomatic characterization of insurance prices [Lien entre stop-loss, TVaR et principes actuariels de tarification]. *Insurance: Mathematics and Economics*, *21*(2), 173–183.
30. Wüthrich, M. V. (2018). *Machine learning in insurance: Theory and case studies*. Springer.
31. Zehnwirth, B. (2002). The negative binomial distribution in insurance. *Insurance: Mathematics and Economics*, *30*(1), 33–47.