Machine Learning Applications in Moroccan Health Insurance Reserving: Neural Network Models and Risk Profiling

Ayyoub SAOUDI[#1], Ghita HAJRAOUI[*2], Jamal ZAHI[#3]

1 University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco.

2 University Hassan II, ESSEM Business School, LARSEMO, Casablanca, Morocco

3 University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco

Email 1 - a.saoudi@uhp.ac.ma

Email 2 - ghita.hajraoui@gmail.com

Email 3 - zahi71@hotmail.com

***Abstract*— We present a comprehensive investigation into neural network-based predictive modeling for health insurance reserve optimization in Morocco. Our methodology integrates advanced machine learning techniques with actuarial risk assessment to develop a hybrid framework for reserve prediction. Decision tree analysis identifies key risk factors like age demographics, chronic disease indicators and gender distributions, that significantly influence claim patterns and reserve requirements.**

The proposed ANN-based approach demonstrates superior predictive performance while the risk profiling component enables policyholder segmentation, providing insurance operators with enhanced tools for financial planning and risk management in the Moroccan healthcare insurance landscape.

***Keywords*—Artificial Neural Networks, Decision Trees, Health Insurance, Modeling techniques, Morocco, Reserving Prediction, Risk Profiling.**

I. INTRODUCTION

The Moroccan health insurance sector is integral to safeguarding the well-being of its citizens by offering crucial healthcare coverage. However, the effective management of this sector heavily relies on the accurate prediction of health insurance reserves. These reserves enable insurance companies to meet their obligations, such as covering medical expenses and compensating policyholders in the event of a claim.

In the past, traditional methods of health insurance reserving prediction have been employed, yet they often fall short in terms of accuracy and efficiency. These methods may rely on simplistic models or historical data trends, which can fail to capture the complex and dynamic nature of healthcare costs and insurance claims. Consequently, inaccuracies in reserve predictions can lead to financial instability within insurance companies, threatening their ability to meet future obligations and potentially undermining the entire healthcare system. [1]

Recognizing these challenges, there is a pressing need to explore advanced techniques that can improve the accuracy and reliability of health insurance reserving prediction. Artificial Neural Networks present a promising solution in this regard. ANN are computational models inspired by the structure and functioning of the human brain, capable of learning complex patterns and relationships from data. By leveraging the power of ANN technology, it becomes possible to analyze vast amounts of data and uncover hidden patterns that may elude traditional statistical methods.[2]

ISSN: 1737-9296

This study aims to unlock the potential of ANN technology to enhance the accuracy of health insurance reserving prediction specifically tailored to the Moroccan context. By utilizing ANNs, we seek to develop predictive models that can better anticipate future healthcare costs and insurance claims, thereby improving the financial stability and sustainability of insurance companies operating within the Moroccan health insurance sector. Through this approach, we aspire to contribute to the effectiveness of the Moroccan healthcare system, ensuring that it remains robust and resilient in the face of evolving healthcare challenges.

II. Health Insurance Reserving

Health insurance plays a critical role in mitigating rising healthcare costs and ensuring equitable access to healthcare services. Accurate reserve estimation is fundamental for insurers to fulfill their obligations to policyholders. Solvency, directly linked to reserving quality, demands increasingly precise and sophisticated methods to guarantee financial stability.

Risk management is another crucial aspect of health insurance. Insurers must anticipate and accurately assess potential risks. Machine learning (ML) methods applied to individual reserving offer a dynamic solution for enhanced risk management. This empowers insurers to safeguard solvency and adapt to evolving trends and changes in policyholder risk profiles. The core challenge lies in optimizing individual reserving to ensure both insurer solvency and policyholder satisfaction.

Traditionally, deterministic methods, such as the Chain Ladder technique [3], have been employed for technical reserve estimation in health insurance. While historically reliable, these methods struggle to capture the complexities of individual characteristics and the dynamic risk landscape. The growing volume of healthcare data renders conventional models inadequate in fully exploiting this information.

Alternative methodologies, like Bornhuetter-Ferguson, Cape Cod, and Benktander-Hovinen, largely build upon the Chain Ladder's core concepts. Run-off triangles aid in comprehensively tracking claim costs, providing valuable insights into their distribution, and facilitating the selection of appropriate modeling techniques. Deterministic methods may suffice for claims with Gaussian distribution, while stochastic methods are more common for modeling other distribution types.[4]

The emergence of ML methods presents an opportunity to address the limitations of traditional techniques in individual reserving for health insurance. Pioneering work by P. Mulquiney [5] introduced neural networks to individual reserving, marking a significant advancement. However, new challenges have arisen, particularly regarding the static or dynamic nature of data. Accounting for the chronological evolution of certain variables becomes crucial for accurate claims payment modeling.

III. Artificial Neural Networks for Enhanced Performance in claim Reserving prediction

Artificial Neural Networks (ANN) draw inspiration from the intricate neural structure of the human brain. [6] In the realm of reserving, they serve as predictive tools, enabling the estimation of diverse outcomes like projected yearly medical claims within insurance firms. ANN possess the ability to learn from past experiences and extrapolate insights, rendering them adept at forecasting forthcoming values using historical data and model outcomes. ANN demonstrates the potential to enhance precision and computational efficiency for forecasting tasks pertinent to reserving, particularly within sectors such as insurance. [5]

The number of epochs and layers in an ANN are important parameters that determine the network's ability to learn from the data and make accurate predictions. The training process involves adjusting these parameters for maximum accuracy.

In the context of ANNs, layers refer to the different levels of interconnected neurons within the network. The simplest type of layer is the input layer, which receives the initial data. Intermediate

ISSN: 1737-9296

layers are known as hidden layers. These layers process the input data through a series of mathematical transformations, extracting and learning features. The output layer produces the final prediction or classification based on the learned features from the hidden layers. The number of layers and the number of neurons within each layer are architectural hyperparameters that influence the network's capacity to learn complex patterns from the data. Deep neural networks contain multiple hidden layers, enabling them to learn hierarchical representations of the input data, whereas shallow networks have fewer hidden layers. [7]

On another hand, an epoch refers to a single pass through the entire training dataset during the training phase of the neural network. During each epoch, the neural network iteratively adjusts its weights and biases based on the error between the predicted output and the actual output, using techniques such as backpropagation and gradient descent. The number of epochs is a hyperparameter that defines how many times the entire dataset will be used to train the neural network. [8]

Optimizing the number of epochs and the architecture of layers in an ANN is essential for achieving optimal performance in terms of accuracy, generalization, and computational efficiency. This optimization process often involves experimentation and tuning of these hyperparameters based on the specific characteristics of the dataset and the requirements of the task at hand.

IV. METHODOLOGY

The estimation of individual reserving in health insurance using Artificial Neural Networks necessitates a systematic and comprehensive approach, commencing with meticulous data collection and extending to a rigorous evaluation of the resultant model's performance.

Our data originates from a health insurance company that has chosen to remain anonymous for confidentiality reasons. The dataset, comprising settlements from the year 2019, covers a variety of variables capturing diverse aspects of insured individuals. These variables contain demographic details such as gender, activity and age, as well as pertinent healthcare-related information, including medical procedure category, predisposition to long-term illness, average cost of medical procedures and the anticipated reimbursement amount for the insured.

To refine the dataset quality and optimize subsequent analyses, several preprocessing steps were meticulously implemented. A crucial aspect of this preprocessing involved the transformation of categorical variables into numerical formats, ensuring compatibility with the ANN model and facilitating effective interpretation and utilization of this information during model development.

The entire process, from data acquisition to preprocessing and the actual implementation of the ANN model, was conducted using the Python programming language. Recognized for its versatility and the availability of comprehensive machine learning libraries, Python facilitated seamless data manipulation, model development, and precise performance assessment.

Additionally, the CART decision tree was generated using RapidMiner, an integrated machine-learning platform. RapidMiner is highly oriented towards code-free development and features a workflow-based machine learning modeling tool coupled with an engine that automates algorithm selection and parameter configuration (auto ML). It encompasses components for hyperparameter tuning and automatic feature engineering, streamlining the model-building process and enhancing efficiency.

Following the model's construction, the reserving amounts were estimated using the ANN model. Subsequently, the health insurance claims were segmented into risk classes based on these estimations. Each segment underwent profiling to analyze specific characteristics, contributing to a thorough understanding of the risk profiles associated to different policyholders.

ISSN: 1737-9296

This methodical approach ensures a dependable and resilient estimation of individual reserving in the domain of health insurance. The resulting model not only encapsulates the inherent complexity of individual characteristics but also underscores the advancements offered by machine learning techniques, specifically ANNs. The insights collected from this process not only furnish valuable information into the complexity of health insurance reserving but also establish a robust groundwork for informed decision-making and strategic initiatives within the broader realm of health insurance management and risk assessment.

V. RESULTS AND DISCUSSION

A. Analysis of Reserving Prediction via ANN Algorithm:

A.1. Model Development and Training:

Initially, the entire dataset is partitioned into two distinct segments. The first segment constitutes the Initial Training Set, comprising 60% of the complete dataset. This subset is dedicated to the construction and preliminary training phases of the ANN model. Concurrently, the remaining 40% of the dataset forms the Temporary Set, which will be subsequently utilized for model validation and testing purposes.

Further division of the Temporary Set involves splitting it into two equal parts. The first subset is designated as the Validation Set, encompassing 20% of the initial data. This subset serves the crucial function of fine-tuning the hyperparameters of the ANN model during the training process, thereby mitigating the risk of overfitting. Meanwhile, the second subset, constituting 20% of the initial data, is denoted as the Test Set. This subset plays a pivotal role in assessing the ultimate performance of the ANN model on unseen data, providing a reliable measure of its predictive capabilities.

This partitioning into three distinct sets is essential to ensure the robustness and generalization of the model. The training set will be utilized to construct the network architecture and learn the underlying relationships within the data. The validation set will enable the optimization of model hyperparameters to prevent overfitting on the training set. Finally, the test set will provide an objective evaluation of the model's ability to generalize its knowledge to unseen data.

A..2. Model Construction and Compilation:

Once the data preprocessing is completed, the next step involves the construction and compilation of the ANN model. During this phase, the architecture of the neural network is defined, including the number of layers, the number of neurons in each layer, and the activation functions used.

A sequential model is instantiated using the Sequential module from Keras. Seven Dense layers are added to the model employing the ReLU activation function. The first layer specifies the input shape of the model based on the number of features in the training data.

The 'adam' optimizer is employed for training the model. The loss function is defined as 'mean_squared_error' to quantify the discrepancy between predicted values and true values. Then, two metrics, 'mean_absolute_error' and 'RootMeanSquaredError', are specified to monitor the model's performance during training.

A.3 Model Results and evaluation metrics:

ISSN: 1737-9296

Our study delves into using ANN to enhance the precision of health insurance reserving prediction within the Moroccan context. The total sum of predictions amounts to 21,037,232 MAD. It represents the value of all predictions made by our model, indicating the overall financial forecast derived from our analysis.

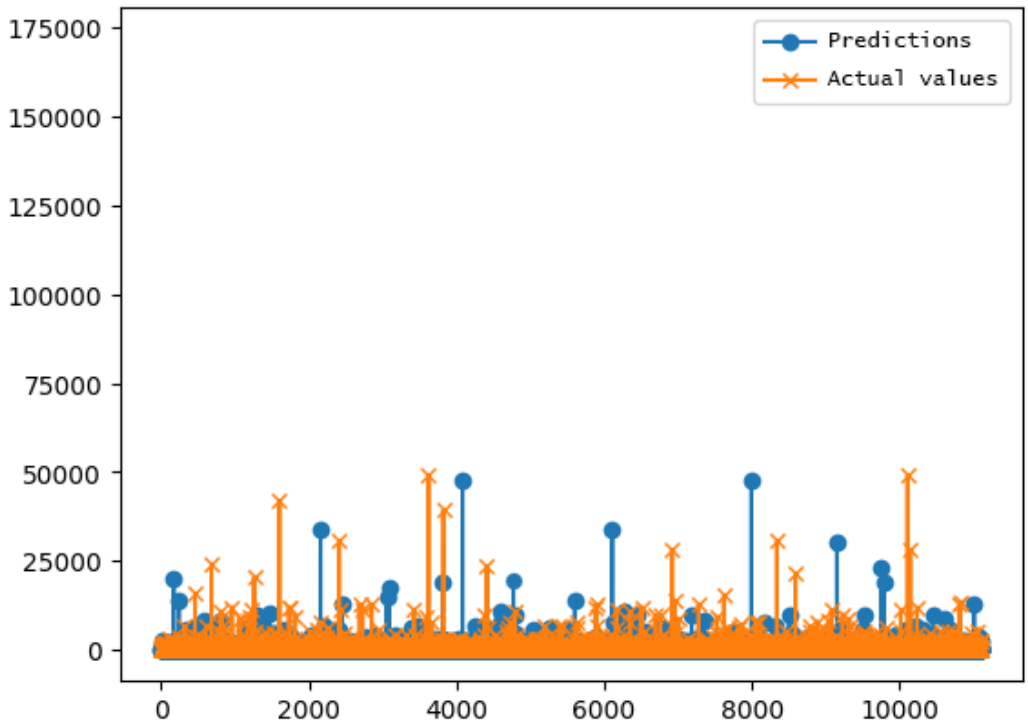

Fig. 1 Comparison between predictions and actual values Source: Designed by the authors

Following the construction and compilation of the ANN model, we can now assess its performance on the unseen data from the Test and Validation Sets. This evaluation is crucial to determine the model's effectiveness in generalizing its learned patterns to new situations. Several metrics are employed to analyze the model's performance:

- Mean Squared Error (MSE): This metric, reported as 682550.75, quantifies the average squared difference between the predicted values and the actual values. A lower MSE signifies a better fit between the predictions and the ground truth.
- Mean Absolute Error (MAE): This metric, reported as 153.55, represents the average absolute difference between the predicted values and the actual values. It provides a simpler interpretation of the average prediction error.
- Coefficient of Determination (R^2): This metric, reported as 86%, indicates the proportion of variance in the target variable explained by the model.

Overall, based on the presented evaluation metrics, it can be concluded that the ANN model built for individual reserving demonstrates satisfactory performance and generalizes effectively to new data.

B. Profiling Policyholders Based on their Risk Level via Decision Trees

ISSN: 1737-9296

Using the results obtained from the provision estimation generated by the ANN model, we proceeded to establish risk profiles for policyholders across three distinct classes: R1, R2, and R3, ordered from least to most risky.

Decision trees were employed as a robust method to delineate these risk classes, leveraging the insights gleaned from the ANN predictions. By partitioning policyholders based on key risk indicators derived from the ANN estimations, decision trees offer a transparent and interpretable framework for profiling policyholders according to their relative risk levels.

This profiling process enables insurance companies to stratify policyholders into risk categories, thereby facilitating targeted risk management strategies and personalized insurance offerings.

B.1 Results of profiling policyholders

```
Activite = actif
|   Risque_Ald = N
|   |   CAT_LIB = BIOLOGIE: R1 {R1=23779, R2=0, R3=0}
|   |   CAT_LIB = DIVERS: R3 {R1=0, R2=0, R3=30}
|   |   CAT_LIB = HOSPITALISATION: R3 {R1=0, R2=0, R3=495}
|   |   CAT_LIB = PHARMACIE: R1 {R1=748, R2=0, R3=0}
|   |   CAT_LIB = RADIOLOGIE
|   |   |   age_cal > 44.500
|   |   |   |   age_cal > 48.500: R3 {R1=0, R2=0, R3=1565}
|   |   |   |   age_cal ≤ 48.500
|   |   |   |   |   ADH_SEXE = Feminin: R3 {R1=0, R2=0, R3=26}
|   |   |   |   |   ADH_SEXE = Masculin
|   |   |   |   |   |   CM > 180
|   |   |   |   |   |   |   age_cal > 47.500: R3 {R1=0, R2=0, R3=3}
|   |   |   |   |   |   |   age_cal ≤ 47.500: R1 {R1=3, R2=0, R3=0}
|   |   |   |   |   |   CM ≤ 180: R1 {R1=12, R2=0, R3=0}
|   |   |   age_cal ≤ 44.500
|   |   |   |   age_cal > 41.500
|   |   |   |   |   MT-Pred_ANN > 1625.157: R3 {R1=0, R2=0, R3=6}
|   |   |   |   |   MT-Pred_ANN ≤ 1625.157
|   |   |   |   |   |   CM > 250
|   |   |   |   |   |   |   age_cal > 42.500: R3 {R1=0, R2=0, R3=3}
|   |   |   |   |   |   |   age_cal ≤ 42.500: R1 {R1=3, R2=0, R3=0}
|   |   |   |   |   |   CM ≤ 250: R1 {R1=21, R2=0, R3=0}
|   |   |   |   age_cal ≤ 41.500: R1 {R1=174, R2=0, R3=0}
|   Risque_Ald = O
|   |   age_cal > 23.500: R3 {R1=3, R2=0, R3=4136}
|   |   age_cal ≤ 23.500: R1 {R1=6, R2=0, R3=0}
Activite = inactif
|   Risque_Ald = N
|   |   age_cal > 46.500: R2 {R1=0, R2=15799, R3=1450}
|   |   age_cal ≤ 46.500
|   |   |   ADH_SEXE = Feminin: R1 {R1=45, R2=0, R3=1}
|   |   |   ADH_SEXE = Masculin
|   |   |   |   age_cal > 41.500: R2 {R1=0, R2=9, R3=0}
|   |   |   |   age_cal ≤ 41.500: R1 {R1=5, R2=0, R3=0}
|   Risque_Ald = O: R3 {R1=0, R2=0, R3=7235}
```

ADH_SEXE	GENDER
CAT_LIB	MEDICAL PROCEDURE CATEGORY
CM	AVERAGE COST
ACTIVITE	POLICYHOLDER ACTIVITY
RISQUE_ALD	PREDISPOSITION TO LONG-TERM ILLNESS
MT-PRED_ANN	PREDICTED REIMBURSEMENT AMOUNT
AGE_CAL	AGE

Fig. 2 Profiling policyholders decision rules Source: Designed by the authors

ISSN: 1737-9296

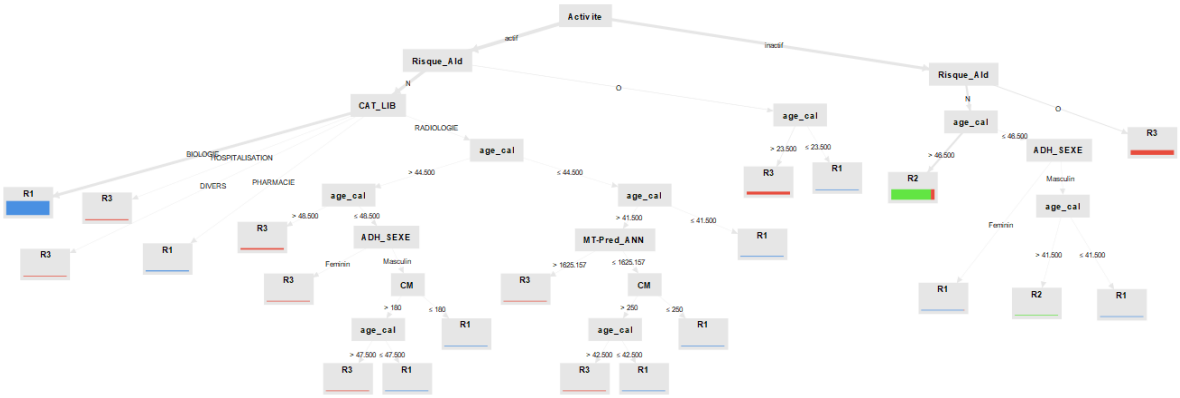

Fig. 3 Decision tree of policyholders' classification Source: Designed by the authors

Based on these decision rules, the profiles of policyholders can be summarized as follows:

TABLE I

CART DECISION RULES

Active Policyholders with No Risk of Ald:
└ For individuals with an activity (actif) and no risk of ALD (Risque_Ald = N), the risk classification (R1, R2, R3) is as follows:
└ Individuals with a medical activity categorized as "Biologie" are classified as R1.
└ Individuals with a medical activity categorized as "Pharmacie" are classified as R1.
└ Individuals with a medical activity categorized as "Divers" or "Hospitalisation" are classified as R3.
└ For individuals with a medical activity categorized as "Radiologie":
└ Those aged over 44.5 years are predominantly classified as R3.
└ Those aged 44.5 years or younger, with a predicted amount (MT-Pred_ANN) over 1625.157 MAD and an average cost (CM) over 250, are classified as R3.
└ Those aged 44.5 years or younger, with a predicted amount (MT-Pred_ANN) over 1625.157 MAD and an average cost (CM) 250 or below, are classified as R1.
└ Those aged 44.5 years or younger, with a predicted amount (MT-Pred_ANN) 1625.157 MAD or below, are classified as R1.
└ Those aged over 41.5 years and 44.5 years or younger, with an average cost (CM) over 250, are classified as R3.
└ Those aged over 41.5 years and 44.5 years or younger, with an average cost (CM) 250 or below, are classified as R1.
└ Those aged 41.5 years or younger are predominantly classified as R1.
Active Policyholders with Risk of Ald:
└ For individuals with an activity (actif) and a risk of ALD (Risque_Ald = O):
└ Those aged over 23.5 years are predominantly classified as R3.
└ Those aged 23.5 years or younger are predominantly classified as R1.

ISSN: 1737-9296

Inactive Policyholders with No Risk of Ald:
└ For individuals with no activity (inactif) and no risk of ALD (Risque_Ald = N):
└ Those aged over 46.5 years are predominantly classified as R2.
└ Those aged 46.5 years or younger:
└ Females are predominantly classified as R1.
└ Males aged over 41.5 years are predominantly classified as R2.
└ Males aged 41.5 years or younger are predominantly classified as R1.
Inactive Policyholders with Risk of Ald:
└ For individuals with no activity (inactif) and a risk of ALD (Risque_Ald = O):
└ Those aged over 46.5 years are predominantly classified as R3.

B.2 Relevant Variables

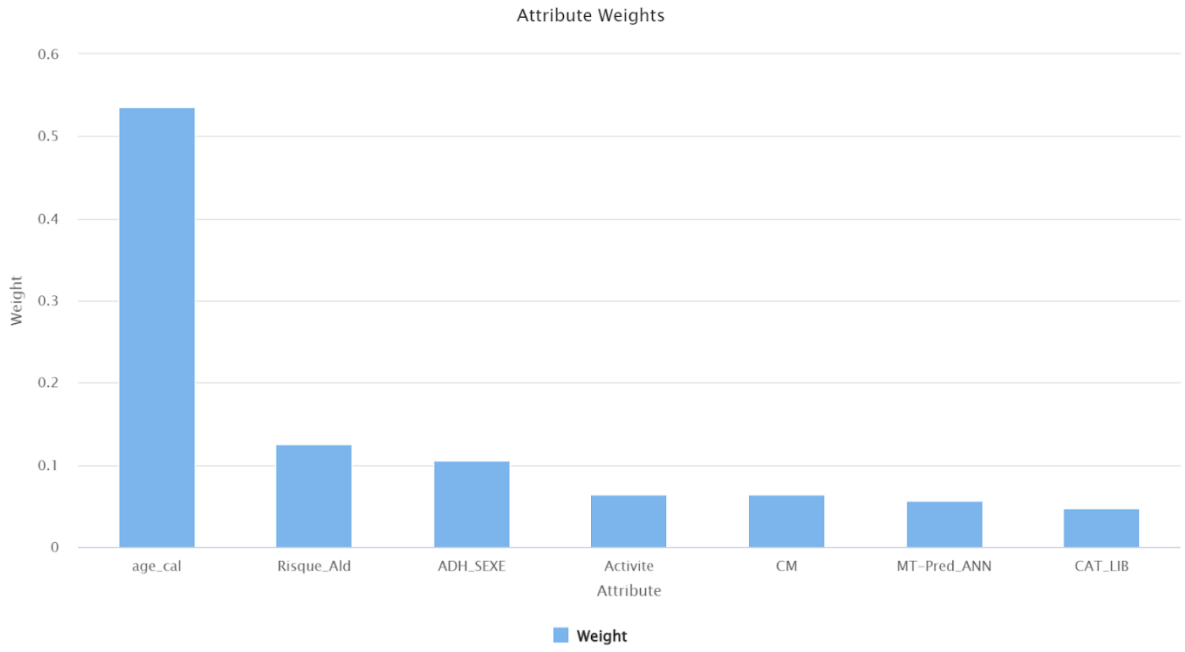

Fig. 4 Classification's relevant variables **SOURCE: DESIGNED BY THE AUTHORS**

Based on the profiling conducted using the CART decision tree algorithm, it becomes evident that the most salient selection variables include the age of the insured individual, wherein risk escalates with advancing age. Subsequently, the presence of a Long-Term Disease (Risque-ALD) emerges as the second variable, signifying that the existence of a chronic ailment amplifies the risk level of the insured individual. Following this, the gender variable is identified, succeeded by the activity status of the insured individual, encompassing all qualitative variables instrumental in determining the individual's risk level. Finally, the average cost, predicted amount, and category of healthcare procedure are integrated as significant determinants.

ISSN: 1737-9296

VI. CONCLUSION

In this study, we explored the use of artificial neural networks to improve the prediction of health insurance reserves in Morocco. Starting with a thorough analysis of relevant variables and a rigorous methodology ranging from data collection to the construction and evaluation of the ANN model, we demonstrated the ability of this approach to provide accurate estimates of insurance reserves. In addition, using decision tree-based risk class profiling techniques, we have identified the main factors influencing policyholders' risk levels, including age, presence of chronic illness and gender.

Our results underline the importance of an integrated approach, combining advanced modeling techniques such as ANNs with risk profiling methods for effective health insurance reserving management. This approach not only provides more accurate reserve estimates, but also a better understanding of the underlying factors that contribute to policyholder risk. This research contributes significantly to the advancement of actuarial science within Morocco's health insurance sector by providing a methodological framework that enhances financial stability through improved reserve accuracy and risk assessment capabilities. The integration of neural network prediction models with comprehensive risk profiling enables insurance companies to optimize capital allocation, reduce uncertainty in reserve estimation and strengthen their solvency position against unexpected claim fluctuations.

Furthermore, the enhanced risk stratification methodology facilitates more precise premium pricing, personalized policy offerings, and targeted risk management strategies, ultimately translating into improved service quality, reduced claim processing times and more equitable coverage options for policyholders across diverse demographic segments.

REFERENCES

[1] LOPEZ, Olivier, MILHAUD, Xavier, et THÉROND, Pierre-E. A tree-based algorithm adapted to microlevel reserving and long development claims. ASTIN Bulletin: The Journal of the IAA, 2019, vol. 49, no 3, p. 741-762.

[2] GOUNDAR, Sam, PRAKASH, Suneet, SADAL, Pranil, et al. Health insurance claim prediction using artificial neural networks. International Journal of System Dynamics Applications (IJSDA), 2020, vol. 9, no 3, p. 40-57.

[3] E. Astesan, "Les réserves techniques des sociétés d'assurances contre les accidents d'automobiles," Librairie générale de droit et de jurisprudence, 1938.

[4] A. Saoudi, F. El Kassimi and J. Zahi, "Technical reserving in non-life insurance: A literature review of aggregated and individual methods, " Journal of Integrated Studies In Economics, Law, Technical Sciences & Communication, Vol (1), No (2) 2023.

[5] P. Mulquiney, "Artificial Neural Networks in Insurance Loss Reserving," in Joint Conference on Information Sciences, 2006.

[6] GROSSI, Enzo et BUSCEMA, Massimo. Introduction to artificial neural networks. European journal of gastroenterology & hepatology, 2007, vol. 19, no 12, p. 1046-1054.

[7] HAYKIN, Simon. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1998.

[8] GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.", 2022.

ISSN: 1737-9296