International Journal of Computer Engineering and Data Science (ISSN: 2737-8543)
Volume 3– Issue 2, April – June 2023

Health insurance risk classification using multinomial logistic regression

Fatima EL KASSIMI[1], Ayyoub SAOUDI[2], Jamal ZAHI[3]

[1] PhD, University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco
[2] PhD student, University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco
[3]Full Professor, University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco

E-mail : f.elkassimi@uhp.ac.ma, a.saoudi@uhp.ac.ma, zahi71@hotmail.com.

Article history

Received Jun 15, 2023
Revised Jun 24, 2023
Accepted Aug 7, 2023
Published Aug 16, 2023

ABSTRACT

In a health insurance portfolio, not all policyholders possess similar risk levels; specific individuals exhibit higher risk than others. Consequently, it might appear inequitable to impose the same premium on everyone. This diversity can be mitigated by employing risk categories that exhibit greater uniformity, considering factors such as gender, age, and other indicators. By applying risk classification, the expected cost for each risk category can be estimated using predefined methods. This study introduces an approach for categorizing insured individuals in health insurance based on statistical learning, specifically employing the multinomial logistic regression algorithm. The research underscores the significance of risk classification in establishing an equitable pricing structure.

Keywords: ***Health insurance, classification, statistical learning, multinomial logistic regression.***

I. INTRODUCTION

The insurance company has a societal responsibility to foster solidarity among its policyholders. By striking a balance between segmenting insureds and pooling risks [1], the insurer prevents adverse selection by offering adequate coverage that involves risk sharing for all policyholders. In line with this objective, this study seeks to refine the pricing structure by establishing four comparable risk categories, ensuring the homogeneity of risks within each class. This approach aims to prevent discriminatory pricing while still segmenting insured individuals. Our methodology involves adjusting the average cost model and subsequently categorizing policyholders based on their risk levels, aligning each policyholder with an appropriate risk class in accordance with the principle of solidarity. To address our specific problem, we will employ the polytomous logistic regression technique as a classification algorithm, enabling us to identify the characteristics of each risk class.

Our article is organized as follows: we present our database and the approach to constructing our target variable in section 2. Section 3 focuses on the polytomous logistic regression model, emphasizing its application principle, selecting the suitable model, and evaluating its quality. Section 4, in turn, presents the results obtained while providing a brief discussion, the model's performance is provided briefly in section 5, and we conclude our work in section 6.

II. DATASET PRESENTATION

Our portfolio is managed by a Moroccan mutual health insurance company operating within the private sector and subject to the National Social Security Fund regulations. Within our database, we have recorded data about 98,000 health insurance claims observed throughout the year 2019. The database comprises 96,540 rows and 20 variables following the necessary data processing steps. Which consist of a *stepwise* variables' selection, only six specific attributes will be utilized as classification features (on a side note the excluded features are not related to our subject of study nor could they be considered for classification purposes). The Table presented below outlines the various features that have the potential to elucidate the claims experience of the insured individuals, encompassing both those who make payments and those who do not. Each entry in the Table pertains to an individual policyholder and encompasses the following characteristics:

TABLE I. FEATURES SELECTION

Feature	Significance	Modalities
Tage	Age range	T1 : [0,10[, T2 : [10,20[, T3 : [20,30[, . . ., T7 : [60,70[, T8 : 70 and plus.
Sex	Gender	M: male, F: female
Cd	Presence of Chronic disease	Y: yes, N: no
BT	Beneficiary type	A: insured themselves; C: insured spouse; E: Insured son or daughter.
CSP	Socioprofessional Catégory	C: single; M: married; D: divorced; V: Widowed.
CAT_LIB	Nature of the care consumed	Cat_lib 1:param_act Cat_lib 2: surgical_act Cat_lib 3: others Cat_lib 4: biology Cat_lib 5: dialysis Cat_lib 6: medical_device Cat_lib 7: med Cat_lib 8: oncology Cat_lib 9: pharmacy_ald. Cat_lib 10:radio

A. *Dependent variable construction*

The endogenous variable we are concerned with is the risk associated with the average cost, referred to as "CM risk" (as we are primarily interested in the expenses incurred by the insured individuals). This variable is categorized into four distinct classes:

TABLE II. RISK CATEGORIES OF INSUREDS

Risk_CM	Risk class	Average cost per procedure	Number of insureds	Final classes
R1	Low risk	<500	90 215	4500
R2	Less risky	≥500,<1000	3 460	3460
R3	Risky	≥1000,<5000	2 535	2535
R4	Highly risky	≥5000,<280001	735	735

A prevalent issue in statistical learning arises when one or more classes dominate the dataset, as is evident in the first class. Such an imbalanced distribution poses a risk of introducing biases in any modeling efforts and may hinder the accurate identification of distinct profiles among individuals belonging to these classes. To address this challenge, we have employed a random under-sampling technique for the overrepresented class, as indicated in the last column of Table I. This approach aims to mitigate the data imbalance and ensure a more equitable representation of the various classes, enabling a more comprehensive analysis of individual profiles.

III. POLYTOMOUS LOGISTIC REGRESSION

Logistic regression is a widely used and well-established classification model that efficiently addresses various classification challenges. It is a valuable tool when the response variable Y is either dichotomous or polytomous, accommodating cases where the explanatory variables encompass qualitative and quantitative attributes.

A. *Logistic regression in health insurance*

Machine learning (ML) in healthcare have improved disease diagnosis and anticipation, enhancing people's lives. Digital health insurance eliminates distance barriers, allowing insurers to offer faster services. Using ML, insurance companies can create efficient policies, efficiently estimate reserves [2], and accurately predict health insurance premiums based on individual features [3] and [4].

Logistic regression (LR) has been frequently employed in health insurance in prior studies to categorize insureds effectively. Previous research, such as [5] conducted a study using data from the Ghana National Health Insurance Scheme revealing that factors such as sex, age, marital status, distance, and length of stay at the hospital significantly influenced health insurance claims. However, health status, billed charges, and income level were not found to be good predictors.

The usage of LR in health insurance has proliferated in recent years, as evidenced by the growing body of literature, including works such as [6] who used data from the Indonesian Family Life Survey and identified factors such as job, education, chronic condition, marital status, and inpatient care as statistically significant predictors of health insurance ownership, while gender and health condition were not significant. Furthermore, their results showed that the probability of having health insurance increased with age. Moreover [7] aimed to identify demographic factors influencing health insurance claim amounts in Saudi Arabia. The logistic model used in their analysis revealed significant factors such as age, gender, nationality, and marital status, which accounted for 90.8% of the variation in health insurance claims. Furthermore [8] compared the Logistic Regression algorithm to The Decision Tree algorithm in classifying cost prediction in health insurance, The results indicate that the Decision Tree algorithm is more effective in predicting and classifying health insurance costs.

Understanding the factors that affect health insurance premiums is crucial for insurance companies to accurately determine their charges. [9] utilized predictive analytics to identify significant factors such as BMI, smoke status, age, and children in charge that impact health insurance costs. Regression and statistical models were employed, and Random Forest emerged as the most effective model, followed by Support Vector Machine. The use of ML in health insurance is not limited to LR, [10] used neural network model to predict health insurance premiums based on personal features such as age, gender, BMI, number of children in charge, smoking habits, and geolocation. their model achieved a high accuracy in predicting health insurance costs.

B. *Model specification*

In our study, the dependent variable "Risk_CM" represents a qualitative response with four distinct categories, necessitating the application of polytomous logistic regression. This involves conducting three separate binomial logistic regressions, each corresponding to one combination of the reference class with the remaining three classes. We designate R1 as the reference class, thus requiring the implementation of

three distinct binomial logistic regression models. (Please note that the we used the notation used in [8] in all of the following formulas).

$$\log\left(\frac{P(Y=R2/X=x)}{P(Y=R1/X=x)}\right)=\log\left(\frac{\pi_2(x)}{\pi_1(x)}\right)=\beta_{02}+\beta_{12}x+\varepsilon \quad (1)$$

$$\log\left(\frac{P(Y=R3/X=x)}{P(Y=R1/X=x)}\right)=\log\left(\frac{\pi_4(x)}{\pi_1(x)}\right)=\beta_{03}+\beta_{13}x+\varepsilon \quad (2)$$

$$\log\left(\frac{P(Y=R4/X=x)}{P(Y=R1/X=x)}\right)=\log\left(\frac{\pi_4(x)}{\pi_1(x)}\right)=\beta_{04}+\beta_{14}x+\varepsilon \quad (3)$$

With our explanatory variables, we note:

$$x=(x_{sex},x_{Tage},x_{PEC\_BT},x_{CSP},x_{CAT\_LIB}) \quad (4)$$

A value of

$$X=(X_{sex},X_{Tage},X_{PEC\_BT},X_{CSP},X_{CAT\_LIB}) \quad (5)$$

The expression of our logistic model is therefore given by:

$$\pi(x)=\frac{\exp\left(\beta_0+\beta_{1sex}+\beta_{2Tage}+\ldots+\beta_{6CAT\_LIB}+\varepsilon\right)}{1+\exp\left(\beta_0+\beta_{1sex}+\beta_{2Tage}+\ldots+\beta_{6CAT\_LIB}+\varepsilon\right)} \quad (6)$$

Estimating this regression model means estimating its parameters β.

$$\log it\left[P\left(Y=R2\middle|x_{sex},\ldots,x_{CAT\_LIB}\right)\right]=\beta_0+\beta_{1sex}+\beta_{2Tage}+\ldots+\beta_{6CAT\_LIB}+\varepsilon \quad (7)$$

Where:

- π(x) Is the model's dependent variable, namely the average cost risk. It corresponds to the risk level and takes the attributes "R1", "R2", "R3," or "R4";
- β_0 represents the constant of the model; it indicates the reference class " R1 ";
- The β_0, β_1, β_2, β_3, β_4, β_5 and β_6 are the parameters of the model that we seek to estimate; these parameters are associated respectively with the sex of the beneficiary, the age of the insured, the risk of chronic disease, the type of beneficiary, his socio-professional category, and the item of care;
- ε represents the error term, which can be interpreted as random noise representing all underlying non-systematic effects that contribute to the measurement error of the model.

To carry out our classification we will be using the *open source R.*

C. *Selection of the logistic model*

To ensure the robustness and reliability of our findings, we implemented a k=5 folds cross-validation technique. This approach allows us to assess the performance of the logistic regression model across multiple configurations. The validation error of the logistic regression is computed as the average of the errors observed during the five different configurations. Additionally, we evaluate the model using the "F1-score" metric, calculated on the fold excluded during each iteration of cross-validation.

In multinomial logistic regression, the number of coefficients that need to be controlled is substantial, making selecting an appropriate model crucial. To address this, we employ the method of nested models to guide our model selection process. In this approach, a Model is considered nested when all of its variables are also present in Model B, meaning that Model B includes Model A. By comparing the performance of nested models, we can assess the incremental impact of adding or removing variables. The outcomes of the nested models are presented in Table 3, providing insights into the performance and effectiveness of different model configurations.

TABLE III. MODEL PERFORMANCE TEST

	Model	Rsd. df	Rsd. Dev	Test	LR stat.	Pr (chi)
1	1	26949	22345.3			
2	sex	26946	22307.3	1 vs 2	38.04	2.7e-08
3	sex + tage	26925	22212.7	2 vs 3	94.57	2.6e-11
4	sex + tage + Cd	26922	21692.4	3 vs 4	520.34	0.0e+00
5	sex + tage + Cd+ cat_lib	26889	14543.7	4 vs 5	7148.6	0.0e+00
6	sex + tage + Cd+ cat_lib + Csp	26877	14519.2	5 vs 6	24.47	1.7e-02
7	sex +tage + Cd+ cat_lib +Csp +Bt	26871	14510.3	6 vs 7	8.86	1.8e-01

D. *Model quality assessment*

1) *Likelihood log*

To assess the effectiveness of our selected model, various indicators were employed. One commonly used metric in evaluating a logistic regression model is the log-likelihood (LL), which measures the overall quality of the model fit to the data. A higher LL value indicates a better fit of the model to the observed data.

Table 3 illustrates the changes in the LL statistic as different variables were added to the model. Notably, the LL value significantly increased from 520.34 (fourth model) to 7,148.69 (fifth model) upon the inclusion of the "CAT_LIB" variable. Subsequently, the LL value decreased to 24.47 for the sixth model. This pattern suggests that the fifth model, with an LL value of 7,148.69, is the most suitable and best-fitting model among the considered alternatives.

2) *Residual deviance*

An alternative method for assessing the regression quality is utilizing the deviance statistic, which aims to minimize its value. In the context of multiple linear regression, [8] considers deviance analogous to the sum of squares of residuals.

When considering nested models, the difference in deviance between two consecutive models helps determine their respective contributions in explaining the underlying model. Models with a significant deviance contribution,

indicated by a low p-value, are deemed acceptable. Ideally, the chosen model should strike a balance between the amount of explained information (deviance) and the model's complexity (number of parameters introduced), favoring parsimony.

Examining Table III, we observe that the residual deviance has notably decreased from 21,692.42 (corresponding to the deviance of the fourth model) to 14,543.72 upon introducing the "CAT_LIB" variable in the fifth model. This reduction in deviance highlights the impact of the care category in explaining the average cost risk. Consequently, the inclusion of this variable signifies its significance in contributing to the understanding of the underlying factors influencing the average cost risk.

3) Chi-square statistic

The outcomes of the chi-square test align with the findings from the log-likelihood and deviance tests, collectively indicating that the most suitable model is the fifth one, which incorporates the following explanatory variables: (sex, Tage, Cd, CAT_LIB). Consequently, at the 95% significance level, we reject the null hypothesis (H0) and accept that these four variables significantly impact the risk level under study. To further evaluate the performance of our model and support these conclusions, we turn our attention to additional parameters. Specifically, we focus on the confusion matrix (Table 6) and the associated metrics, which will be discussed in the subsequent analysis.

E. Selection Criteria for explanatory variables

Examining the p-value statistic associated with the estimated coefficients in the saturated model is customary, encompassing all potential variables. This analysis helps identify variables that can be excluded from the final model. The results of this test confirm the earlier findings derived from the log-likelihood, deviance, and chi-square tests discussed earlier. Consequently, the collective results from these tests provide substantial evidence regarding the influence of variables such as insured individuals' gender, age, chronic disease, and the specific healthcare service utilized to classify insured individuals according to their risk levels.

Therefore, based on the results obtained, we reject the null hypothesis at the 95% confidence level, affirming that these variables significantly influence our dependent variable.

F. Results and Discussion

The logistic model after the removal of the non-significant variables is as follows:

TABLE IV. ESTIMATION OF THE COEFFICIENTS OF THE MULTINOMIAL REGRESSION MODEL

	Dependent variable		
	R2 (1)	***R3 (2)***	***R4 (3)***
SexMale	1.070 (0.067)	1.362*** (0.077)	1.443*** (0.129)
Tage2	1.253 (0.402)	1.973 (0.437)	9.829*** (0.831)
Tage3	1.624 (0.377)	2.288** (0.418)	55.113*** (0.788)
Tage4	1.019 (0.382)	1.360 (0.423)	13.520 (0.803)
Tage5	0.907 (0.364)	1.512 (0.401)	20.334*** (0.771)
Tage6	1.018 (0.345)	1.556 (0.387)	17.561*** (0.748)
Tage7	0.982 (0.348)	1.148 (0.387)	13.445*** (0.755)
Tage8	1.008 (0.353)	1.333 (0.393)	15.615*** (0.758)
CAT_LIB 1	0.002*** (1.450)	0.00*** (0.00)	0.00*** (0.00)
CAT_LIB 2	59,475*** (0.433)	630,099*** (0.375)	142,487*** (0.685)
CAT_LIB 3	185,286 *** (0.385)	25,310.440*** (0.424)	11,184*** (0.690)
CAT_LIB 4	0.002*** (1.011)	0.001*** (1.009)	0.004*** (1.447)
CAT_LIB 5	308,163*** (0.510)	1,565.370*** (0.510)	0.022*** (0.000)
CAT_LIB 6	0.097** (1.019)	0.062*** (1.016)	0.418 (1.434)
CAT_LIB 7	0.092** (1.057)	0.122** (1.052)	1.685 (1.463)
CAT_LIB 8	0.048*** (0.000)	90,844,173*** (0.508)	4,261,560*** (0.508)
CAT_LIB 9	2,654.36*** (0.986)	1,369.59*** (0.949)	155,904*** (1.207)
CAT_LIB 10	0.087** (1.014)	0.077** (1.011)	0.273 (1.428)
Cd	0.00001*** (0.980)	0.00000*** (0.946)	0.00000*** (0.891)
intercept	52.386*** (1.067)	43.844*** (1.078)	0.047* (1.603)
Akaike inf. Crit.	14,669.720	14,669.720	14,669.720
Note :	* p < 0.1 ;	** p < 0.05 ;	*** p < 0.01

As shown in TABLE VI., the coefficients are not directly interpretable. It is the ODDS ratios (cf. Table V) that we will take into account.

TABLE V. ODDS RATIOS OF THE MULTINOMIAL LOGISTIC MODEL

	Dependent variable		
	R2 (1)	*R3 (2)*	*R4 (3)*
SexMale	0.068 (0.067)	0.309*** (0.077)	0.367*** (0.129)
Tage2	0.226 (0.402)	0.679 (0.437)	2.285*** (0.831)
Tage3	0.485 (0.377)	0.828** (0.418)	4.009*** (0.788)
Tage4	0.018 (0.382)	0.307 (0.423)	2.604 (0.803)
Tage5	-0.098 (0.364)	0.413 (0.401)	3.012*** (0.771)
Tage6	0.018 (0.345)	0.442 (0.387)	2.866*** (0.748)
Tage7	-0.018 (0.348)	0.138 (0.387)	2.599*** (0.755)
Tage8	0.008 (0.353)	0.138 (0.393)	2.748*** (0.758)
CAT_LIB 1	-6.487*** (1.450)	-28.891*** (0.00)	-29.652*** (0.00)
CAT_LIB 2	17.901*** (0.433)	20.261*** (0.375)	25.683*** (0.685)
CAT_LIB 3	12.130*** (0.385)	10.139*** (0.424)	16.230*** (0.690)
CAT_LIB 4	-6.219*** (1.011)	-6.902*** (1.009)	-5.526*** (1.447)
CAT_LIB 5	12.638*** (0.510)	7.356*** (0.510)	-3.795*** (0.000)
CAT_LIB 6	-2.337** (1.019)	-2.773*** (1.016)	-0.872 (1.434)
CAT_LIB 7	-2.388** (1.057)	-2.107** (1.052)	0.522 (1.463)
CAT_LIB 8	-3.041*** (0.000)	18.325*** (0.508)	22.173*** (0.508)
CAT_LIB 9	7,884*** (0.986)	7.222*** (0.949)	11,957*** (1.207)
CAT_LIB 10	-2.440** (1.014)	-2.567** (1.011)	-1.297 (1.428)
Cd	-11.616*** (0.980)	-13.096*** (0.946)	-13.961*** (0.891)
intercept	3.959*** (1.067)	3.781*** (1.078)	-3.057* (1.603)
Akaike inf. Crit.	14,669.720	14,669.720	14,669.720
Note:	* p < 0.1 ;	** p < 0.05 ;	*** p < 0.01

These two tables show that the classification variables have a statistically significant impact on the risk classes. The results of the logistic model state that:

- All other things being equal, male insureds have a 36% chance of being in the fourth class compared to the first class.
- All other things being equal, beneficiaries aged between 10 and 20 are twice as likely to belong to the fourth R4 class than the first.
- *Ceteris paribus*, a person in their thirties is four times more likely to belong to the fourth risk class than the first.
- *Ceteris paribus*, being in one's forties, increases the chances of belonging to the fourth class by two and a half times compared to the first class.
- All other things being equal, beneficiaries aged between 50 and 60 are almost three times more likely to belong to the fourth class than the first class (the same observation is valid for beneficiaries in age groups 7 and 8).
- All other things being equal, insureds consuming paramedical acts are six times less likely to belong to the second class than our reference class R1.
- All other things being equal, consuming a surgical procedure makes insureds 25 times more likely to belong to the fourth class than the first class.
- All other things being equal, consuming a biological procedure reduces the risk of belonging to the fourth class by five times compared to R1.
- All other things being equal, insureds consuming dialysis procedures are much more likely to belong to R2 than our reference class R1.
- Consuming medical devices decreases the chance of belonging to class R4, compared to our reference class R1.
- Consuming medication decreases the probability of belonging to the second and third classes and increases the probability of belonging to R4, compared to our reference class R1.
- Cancer increases the probability of belonging to class R4 by almost 22 times, and to the third class by 18 times, and decreases the chance of belonging to the second class, compared to our reference class R1.
- Consuming long-term care drugs reduce the chance of belonging to the second class by 11 times, compared to R1.
- Consuming a radiology procedure makes you less likely to belong to the fourth class than our reference class R1.
- A chronic disease reduces the chance of belonging to the fourth class 11 times, compared to R1.

G. Model performance

Generally, good predictions are positioned on the diagonal. Thus, here we have:

TABLE VI. CONFUSION MATRIX

	R1	R2	R3	R4
R1	664	190	46	0
R2	73	423	194	2
R3	51	149	285	22
R4	3	39	30	75

Out of 791 insured individuals, the logistic regression model successfully classified 664 individuals as belonging to class R1, representing a reasonably satisfactory result. Similarly, for class R2, the model accurately identified 423 out of 801 insured individuals. In the case of class R3, 285 out of 555 individuals were correctly classified. Class R4 saw the correct identification of 75 out of 99 insured individuals by the logistic regression model. Combining these correct classifications, the total number of True Positives (TPs) amounts to 1447, representing the sum of the values along the diagonal.

The table below resume the evaluation metrics per class:

TABLE VII. CONFUSION MATRIX METRICS

Class	Accuracy	Precision	Recall	F1-Score
1	83.84%	0.73	0.84	0.79
2	71.19%	0.61	0.53	0.57
3	78.09%	0.56	0.51	0.54
4	95.73%	0.51	0.76	0.61
Overall	64.43%	-	-	0.63

Based on the given metrics of 64.43% accuracy and an F1-score of 0.63, we can evaluate the performance of the model as follows:

- Overall Accuracy: An accuracy of 64.43% indicates that the model correctly classified approximately 64.43% of the instances in the dataset. While accuracy is a commonly used metric, it may not always provide a complete picture of model performance in the presence of imbalanced datasets which is not the case in our model. That is why we opted for the F1-score measure.
- F1-score: The F1-score of 0.63 indicates a reasonably balanced performance between precision and recall. It considers both false positives and false negatives, providing a more comprehensive evaluation of the model's effectiveness.

From a statistical perspective, these scores are deemed broadly acceptable, and the multinomial regression classification model appears to have achieved a good level of performance.

IV. CONCLUSION

The multinomial logistic regression algorithm was employed to profile and classify the insured individuals within the portfolio based on their risk classes; this regression technique facilitated the classification of insured individuals by leveraging their unique characteristics, enabling them to be assigned to the appropriate class based on their level of risk.

These outcomes have two significant implications. Firstly, they enable the segmentation of policyholders according to their risk levels, providing insights into the specific profiles associated with each risk class. Secondly, this information facilitates the alignment of each risk class with an equivalent tariff class. By doing so, insured individuals within the same risk class are charged the same premium, thereby promoting fairness in tariff structures. However, it is important to note that a certain degree of mutualization, such as a small percentage, may be considered between these classes to address any disparities.

Overall, the application of the multinomial logistic regression algorithm has proven instrumental in achieving the profiling, classification, and subsequent equitable tariff assignment of insured individuals within the portfolio based on their risk characteristics.

REFERENCES

[1] Charpentier, A., Denuit, M., & Elie, R. (2015). Segmentation et mutualisation les deux faces d'une même pièce. *Risques n° 103*, 19-23.

[2] *SAOUDI,A.,EL KASSIMI, F., ZAHI, J. (2023).*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) 2023,1-12.

[3] EL KASSIMI. F & ZAHI. J (2022). Proposition d'un modèle de tarification en assurance maladie obligatoire à travers le modèle linéaire généralisé . Alternatives managériales et économiques Vol 4, No 4(Octobre, 2022) 462-481.

[4] EL KASSIMI. F & ZAHI. J (2022). Health insurance pricing using CART decision trees algorithm . International Journal of Computer Engineering and Data Science (IJCEDS), 2(3).

[5] Antwi, S., & Zhao, X. (2012). A logistic regression model for Ghana National Health Insurance claims. *International Journal of Business and Social Research, LAR Center Press,vol. 2(7)*, 139-147.

[6] Astari, D. W., & Kismiantini. (2019). Analysis of Factors Affecting the Health Insurance Ownership with Binary Logistic Regression Model. *Journal of Physics : Conf. Series 1320*, doi:10.1088/1742-6596/1320/1/012011.

[7] Osman, M. A., & Ismail, E. A. (2018). A Quantitative Model to Identify Key Determinants for Health Insurance Claims in the Kingdom of Saudi Arabia. *J. King Saud Univ., Vol. 27*, 23-36.

[8] Mounika, K. G., & Deepa, N. (2023). By contrasting decision trees with logistic regression, a novel categorization-based cost prediction method for health insurance may be developed under supervision. *Journal of Survey in Fisheries Sciences*, 1468-1477.

[9] Jun, J. S. (2020). Identification and Prediction of Factors Impact America Health Insurance Premium. Dublin: Masters thesis, National College of Ireland.

[10] Kaushik, K., Bhardwaj, A., Dhar Dwivedi, A., & Singh, R. (2022). Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. *Int. J. Environ. Res. Public Health*. doi:doi:org/10.3390/19137898.

[11] Rakotomalala, R. (2011). *Pratique de la Régression Logistique.* Université Lumière Lyon 2.