SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Health Insurance Premium Pricing Using Machine Learning Methods

Ahmad Nur Azam Ahmad Ridzuan[1,*], Aina Zafirah Azman[1], Fatin Alya Marzuki[1], Wan Shazmien Danieal Mohamed Faudzi[1], Siti Hajar Abd Aziz[2], Norida Abu Bakar[3]

[1] Department of Actuarial Science, College of Computing, Informatic and Mathematics, UiTM Perak Branch Tapah Campus, 34500 Tapah, Perak, Malaysia
[2] Department of Media Studies, Faculty of Mass Communication, UiTM Melaka Branch Alor Gajah Campus, 78000 Alor Gajah, Melaka, Malaysia
[3] Department of Business Studies, Faculty of Business and Management, UiTM Melaka Branch Alor Gajah Campus, 78000 Alor Gajah, Melaka, Malaysia

ARTICLE INFO

Article history:
Received 22 June 2023
Received in revised form 28 October 2023
Accepted 2 November 2023
Available online 4 March 2024

Keywords:

Health insurance; Decision tree; Neural network; Healthcare; Premium pricing

ABSTRACT

Health insurance is important alongside life insurance products. This product's subscription is gradually rising and has become one of the public's main considerations due to their awareness on medical and surgical costs. This study is aimed at (i) identifying whether the independent factors are important in predicting the dependent variable, and (ii) to determine which model (logistic regression model/decision trees/neural networks) is the best model to be utilised. The SAS Enterprise Miner (E-Miner) was used to analyse the data and to select the best model. At the end of this study, the measurements like the Maximum Absolute Error (MAE), Average Squared Error (ASE), Root Average Squared Error (RASE), and Sum Squared Error (SSE) in the decision tree model indicated the lowest errors and led to the selection of the decision tree as the best model to be used in health insurance premium pricing.

1. Introduction

Health insurance is crucial because it enables people to receive a timely medical care, which enhances their quality of life and health. Some might think that since emergency rooms are always open, and everyone has access to healthcare. However, even in places where the safety net is strong, it does not remove the obstacles to access the same room in the degree that the health insurance offers. In a recent multiyear evaluation, the Institute of Medicine (IOM) concluded that "coverage matters." When a company and a customer enter a contract for a health insurance, in exchange for the payment of monthly premium, the corporation offers covering all or parts of the insured person's medical expenses. The contract, which typically covers a year, outlines the precise costs linked to disease, injury, pregnancy, or preventative treatment that the insurance will be liable for covering. Each insured person is obligated to pay an agreed-upon premium to the insurance provider for their

* *Corresponding author.*
E-mail address: ahmad558@uitm.edu.my

https://doi.org/10.37934/araset.41.1.134141

medical insurance. Insurance premiums are not fixed in quantity, yet they are determined by several characteristics from the person seeking health and welfare protection.

2. Research Background

The health insurance policy has become an important coverage that needs to be subscribed by the public. In some countries like Indonesia, Ghana, and Rwanda, health insurance coverage contributes in between 3 to 11 percent to these nations' income, whereas some other nations like Gabon and China have shown some growing needs of this policy [3,13]. In China, the term "Healthy China 2030" envisions the Chinese citizens to get a full health protection [21]. The health care utilization produces a gap not only towards the rural and urban people, but it also affects those who belong under the low and high-income earners [11]. In Malaysia, [12] indicated that age is considered as the biggest factor affecting Malaysians' decision to purchase a health insurance policy [12], also concluded that salary and willingness to pay for the insurance policy are parts of the factors affecting the decision to buy a health insurance policy. [20] indicated that the citizens mostly prefer to discover more about public health policies instead of health care policy approaches [8], found that by introducing the theory of "Getting to know" to American Indian elders, this made them feel comfortable about healthcare system provided by the Health Care Providers (HCPs). [5] concluded that the health policy established under the National Health Promotion Policy (NHPP) in South Africa was inequal, especially to the people who stay in rural areas, as well as the less-educated groups. [7] revealed that health policy plays a major role especially when it is related to the arts and public health, where art activities might influence health policy decisions. [10] indicated that Pakistan has tremendous issues at providing a better health system for its people, especially the absence of a national health policy including the monitoring and evaluating of health programs.

Decision tree is one of the popular instruments that is widely used as a machine learning technique. [18] used the decision tree method to examine the gaps in minor health disorders in Switzerland. [15] found that machine learning models like linear regression, decision tree, and artificial neural network (ANN) helped them improve budget allocation in public-funded insurance sectors. [14] found that the introduction of machine learning techniques in the insurance industry was truly helpful, as it helped them to understand the patterns easily by using the models. [16] used tree diagrams and were able to compare the needs for primary care (PHC), and advanced care (AHC), to determine cancer care coverage rates in Indonesia. [19] used decision tree as one of the models to investigate diabetes mellitus disease, which is considered as a major health issue in India. [4] used machine learning models like the decision tree, neural network, random forest, support vector machines, and others to predict health conditions, which are applied by the US Health and Retirement Study data. [1] implemented machine learning instruments to detect and prevent fraud in healthcare, specifically during the claim processing. [2] used machine learning algorithms at predicting the likelihood of diabetic diseases where it was found that these algorithms were very helpful at modelling the disease. [6] found that machine learning is not only helpful on predicting healthcare outcomes, but it also guides the health service researchers on generalizing data-driven estimators. [9] used four machine learning models likes the decision tree, random forest, support vector machine, and XGBoost to predict the length of stay (LOS) of patients that are admitted into the South Korean medical institutions and discovered that XGBoost is the best model to be used in this matter.

3. Methodology

The objectives of this paper are to investigate the different features to observe their relationship, plot a multiple linear regression, and access the model assessment from a dataset based on several features of individuals such as age, physical or family condition, and location against their existing medical expenses for the purpose of predicting future medical expenses of individuals, to assist medical insurance companies determine their premiums.

3.1 Data Description

The dataset was originated from Brett Lantz's book "Machine Learning With R". The dataset is now accessible online via the "Machine Learning With R" repository on GitHub. 1338 observations (rows) and 7 features make up the dataset (columns). Four numerical features (age, BMI, children, and charges), three nominal features (sex, smoker, and region) were translated into factors with numerical values assigned to each level. For this research, the SAS Enterprise Miner (E-Miner) was used as an instrument to clean and analyse the dataset. Table 1 shows the dataset used in the study. The data explains the variables used in this study together with the descriptions and data types that were applied in the SAS E-Miner.

Table 1
Descriptions of Dataset

Column name	Description	Data type
age	Primary beneficiary's age	Integer
sex	Beneficiary's gender (male or female)	Character
BMI	Body Mass Index, providing an understanding of body weights that are relatively high or low relative to height, ratio of height to weight	Numeric
children	Number of children covered by health insurance/Number of dependents	Integer
smoker	Whether a person is a smoker (yes) or not (no)	Character
region	Beneficiary's residential area in the United States (northeast, southeast, southwest, northwest)	Character
charges	Individual insurance premiums billed by the health insurance	Numeric

3.2 Method Used

The models used in this study were the decision tree model, and the neural network model. The decision trees produced a set of rules that can be used to generate predictions for a new dataset. This information can then be used to drive business decisions. One advantage of the decision tree is that it produces an output that describes the scoring model with interpretable node rules, whereas the neural network is a series of algorithms that helps to recognize the underlying correlations in a set of data by emulating the way the human brain works. The neural network node provides a variety of feedforward networks that are commonly called as backpropagation or backprop networks. [17] mentioned that neural network can learn new data in making predictions. Basically, this study also identified the lowest error indicated in both models, whereby at the end of the process, it can be used to decide whichever model to be considered as the best model.

4. Results

Figure 1 shows the variables, namely the charges that were selected as the target variable. Here, the SAS E-Miner required the selection of one variable as the target variable.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
age	**Input**	**Interval**	**No**		**No**	.	.
bmi	**Input**	**Interval**	**No**		**No**	.	.
charges	**Target**	**Interval**	**No**		**No**	.	.
children	**Input**	**Ordinal**	**No**		**No**	.	.
region	**Input**	**Nominal**	**No**		**No**	.	.
sex	**Input**	**Nominal**	**No**		**No**	.	.
smoker	**Input**	**Binary**	**No**		**No**	.	.

Fig. 1. Variable Roles and Level

Figures 2, 3 and 4 describe the first model that was used to analyse the dataset. The model is the decision tree. In this study, 10 rules were found, represented by the number of leaves. The root node for this study under the decision tree is SMOKER. The splitting variables are smoker status, BMI, age, and children, and they appeared 1, 2, and 6 times per splitting respectively. On the other hand, sex and region were not spitted in this study as it is considered as an unrelated variable by the SAS.

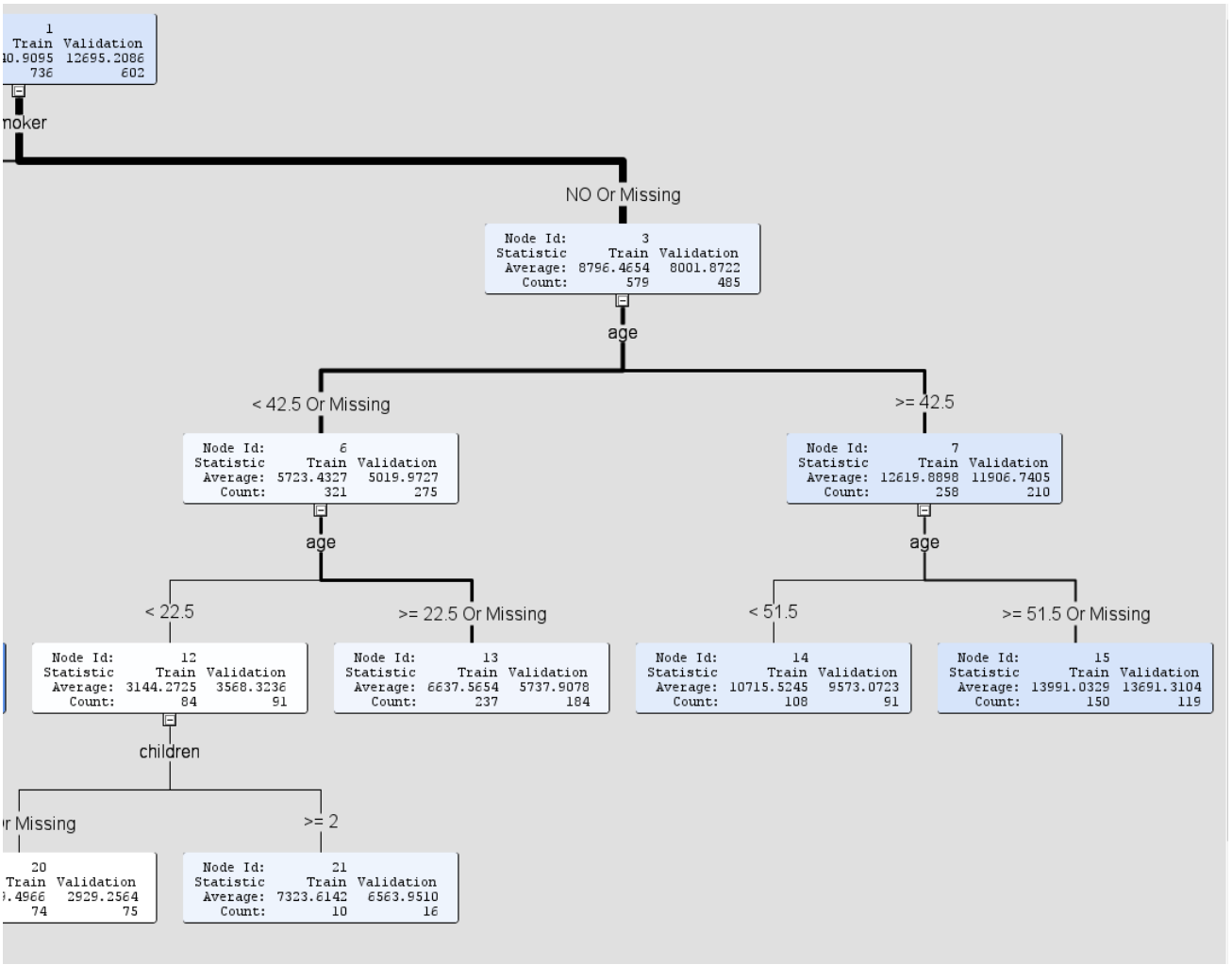

Fig. 2. Decision Tree Model

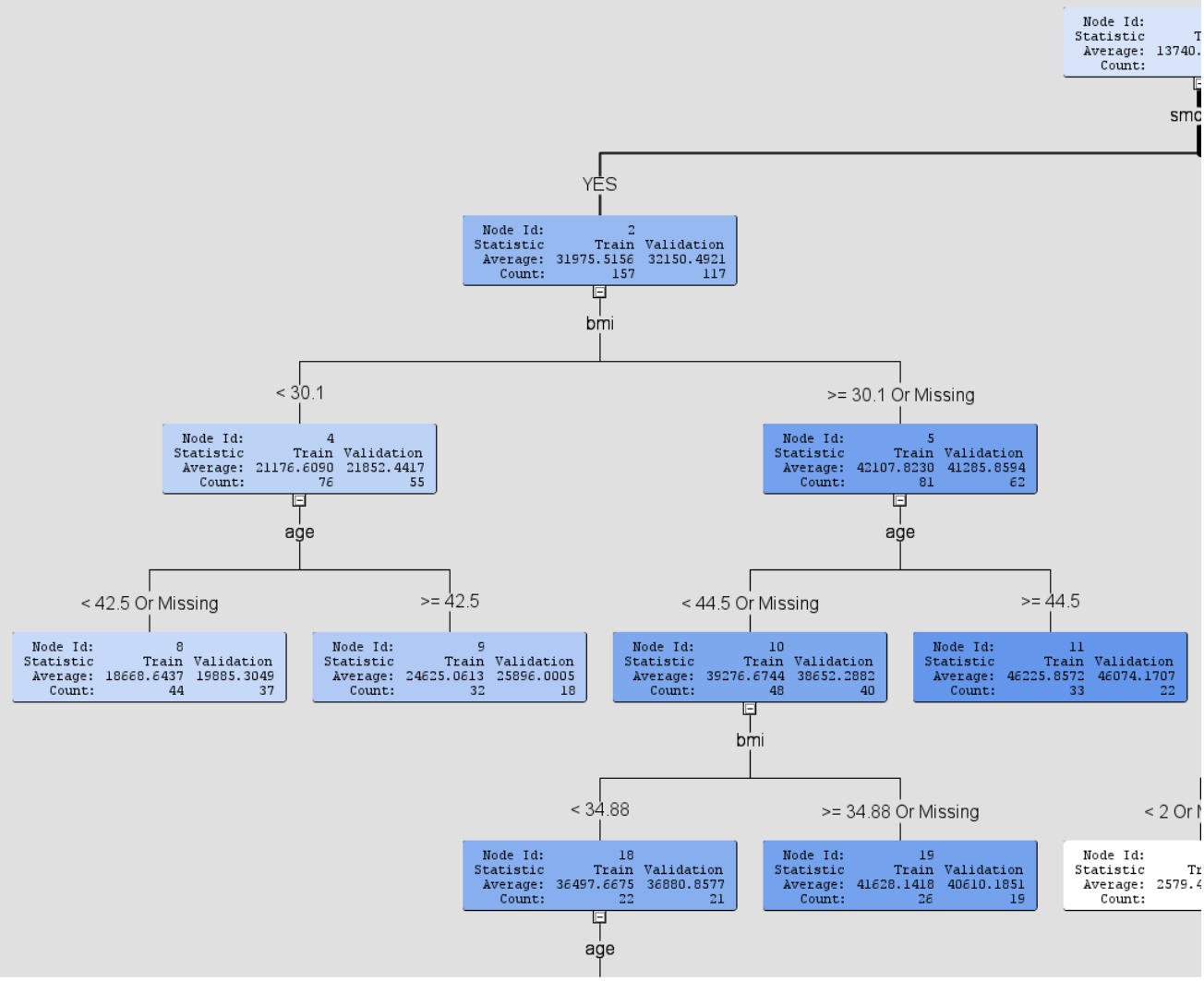

Fig. 3. Decision Tree Model

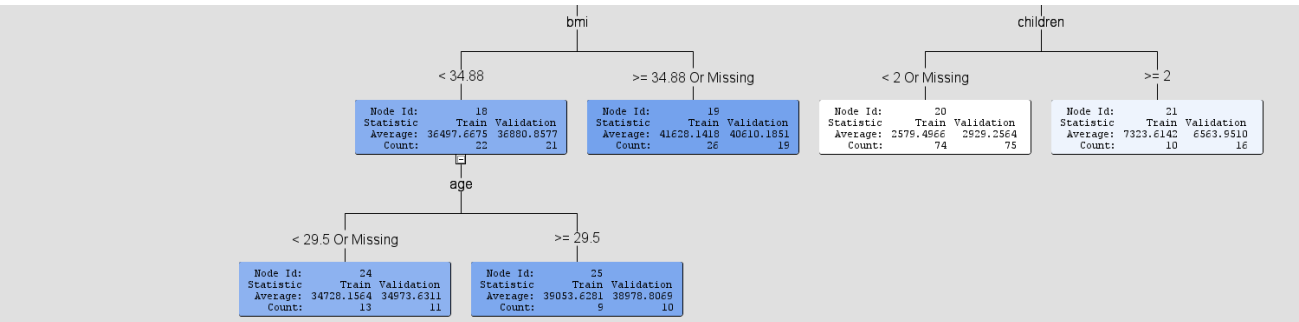

Fig. 4. Decision Tree Model

The second model used for this study was the Neural Network Model. Figure 5 shows the result that was analysed by the SAS using the neural network model. It indicated that the variable of charges was analysed by various statistical measurements. All these tests were used in the neural network model to indicate the model that can be fitted from the data, as well as to look for the errors that exist from the analysis.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
charges		DFT	Total Degrees...	736	.	.
charges		DFE	Degrees of Fr...	671	.	.
charges		DFM	Model Degree...	65	.	.
charges		NW	Number of Est...	65	.	.
charges		AIC	Akaike's Infor...	12590.97	.	.
charges		SBC	Schwarz's Bay...	12890.05	.	.
charges		ASE	Average Squa...	22537017	17214758	.
charges		MAX	Maximum Abs...	21882.97	25185.62	.
charges		DIV	Divisor for ASE	736	602	.
charges		NOBS	Sum of Frequ...	736	602	.
charges		RASE	Root Average ...	4747.317	4149.067	.
charges		SSE	Sum of Squar...	1.659E10	1.036E10	.
charges		SUMW	Sum of Case ...	736	602	.
charges		FPE	Final Predictio...	26903354	.	.
charges		MSE	Mean Square...	24720186	17214758	.
charges		RFPE	Root Final Pre...	5186.844	.	.
charges		RMSE	Root Mean Sq...	4971.94	4149.067	.
charges		AVERR	Average Error ...	22537017	17214758	.
charges		ERR	Error Function	1.659E10	1.036E10	.
charges		MISC	Misclassificati...	.	.	.
charges		WRONG	Number of Wr...	.	.	.

Fig. 5. Neural Network Results

The third analysis implemented in this study was the model comparison. In order to decide whether the Decision Tree or the Neural Network as the best model for this dataset's analysis, this study used the Model Comparison node in SAS Enterprise Miner. The Model Comparison node belongs to the Assess category in the SAS data mining process of Sample, Explore, Modify, Model, and Assess (SEMMA). The Model Comparison node enables the comparison of performance of competing models using various benchmarking criteria.

Table 2 shows the results that were analysed by the SAS. The result was compared between the models used, namely the decision tree and neural network. The shown result compares the errors indicated in the analysis. As it appears in the table, for the Maximum Absolute Error (MAE), the decision tree model was much better compared to the neural network model. The result also shows that the Average Squared Error (ASE), Root Average Squared Error (RASE) as well as Sum Squared Error (SSE) indicates the same result whereby all measurements showed that the decision tree model is much better compared to the neural network model.

Table 2
Comparison Model Result

Model	MAE	ASE	RASE	SSE
Decision Tree	25144.79	16989133	4121.788	1.023E10
Neural Network	25185.62	17214758	4149.067	1.036E10

Therefore, it can be concluded that the decision tree model is the best model to analyse health insurance premium charges. Figure 6 shows the model comparison between the decision tree and the neural network.

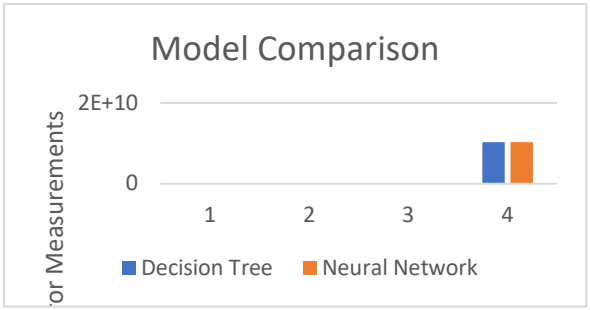

Fig. 6. Model Comparison

5. Conclusions

This study examined the model assessment that can be best suited at measuring health insurance premium charges. The three main objectives were conducted, where the researchers identified whether the independent factors were important at predicting the dependent variable. Secondly, whether the logistic regression model, the decision trees, or the neural networks was the best model to be utilised was also determined.

The SAS's decision tree implementation identified that the root node for this model is SMOKER. In the decision tree results, the number of leaves represents the presence of ten regulations and six significant variables that are prioritised by the value of the "importance" column. Smoker status, BMI, age, and children are the splitting variables, and it appeared 1, 2, and 6 times respectively in each splitting. The second model used in this study was the neural network model. The neural network node provides a variety of feedforward networks, sometimes known as backpropagation or backprop networks. To answer the third objective, this study compared its result from both models (decision tree and neural network) to determine which model can be indicated as the best model. The result is the SAS E-Miner stating the decision tree model as best model because it produced the lowest error. It can be concluded that to measure the health insurance premium charges, the decision tree model should be used as the instrument.

Acknowledgement
This research was not funded by any grant.

References

[1] Amponsah, Anokye Acheampong, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. "A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology." *Decision Analytics Journal* 4 (2022): 100122. https://doi.org/10.1016/j.dajour.2022.100122

[2] Chang, Victor, Meghana Ashok Ganatra, Karl Hall, Lewis Golightly, and Qianwen Ariel Xu. "An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators." *Healthcare Analytics* 2 (2022): 100118. https://doi.org/10.1016/j.health.2022.100118

[3] Choi, Weng I., Honghao Shi, Ying Bian, and Hao Hu. "Development of commercial health insurance in China: a systematic literature review." *BioMed Research International* 2018 (2018). https://doi.org/10.1155/2018/3163746

[4] Chowdhury, Rafiqul I., and Jabed H. Tomal. "Risk prediction for repeated measures health outcomes: A divide and recombine framework." *Informatics in Medicine Unlocked* 28 (2022): 100847. https://doi.org/10.1016/j.imu.2022.100847

[5] Mostert, Cyprian Mcwayizeni. "The impact of national health promotion policy on stillbirth and maternal mortality in South Africa." *Public Health* 198 (2021): 118-122. https://doi.org/10.1016/j.puhe.2021.07.009

[6] Doupe, Patrick, James Faghmous, and Sanjay Basu. "Machine learning for health services researchers." *Value in Health* 22, no. 7 (2019): 808-815. https://doi.org/10.1016/j.jval.2019.02.012

[7] Dow, Rosie, Katey Warran, Pilar Letrondo, and Daisy Fancourt. "The arts in public health policy: progress and opportunities." *The Lancet Public Health* 8, no. 2 (2023): e155-e160. https://doi.org/10.1016/S2468-2667(22)00313-9

[8] Haozous, Emily A., Elise Trott Jaramillo, and Cathleen E. Willging. "Getting to know: American Indian Elder health seeking in an under-funded healthcare system." *SSM-Qualitative Research in Health* 1 (2021): 100009. https://doi.org/10.1016/j.ssmqr.2021.100009

[9] An, Jiho, Mungyo Jung, Seiyoung Ryu, Yeongah Choi, and Jaekyeong Kim. "Analysis of length of stay for patients admitted to Korean hospitals based on the Korean National Health Insurance Service Database." *Informatics in Medicine Unlocked* 37 (2023): 101178. https://doi.org/10.1016/j.imu.2023.101178

[10] Khan, Mohsin Saeed, and Babar Tasneem Shaikh. "What does it take to make a wrong decision? A qualitative study from Pakistan's health sector." *Dialogues in Health* 2 (2023): 100127. https://doi.org/10.1016/j.dialog.2023.100127

[11] Li, Chaofan, Chengxiang Tang, and Haipeng Wang. "Effects of health insurance integration on health care utilization and its equity among the mid-aged and elderly: evidence from China." *International Journal for Equity in Health* 18 (2019): 1-12. https://doi.org/10.1186/s12939-019-1068-1

[12] Selamat, Ellyana Mohamad, Siti Rasidah Abd Ghani, Nurcholisah Fitra, and Faiz Daud. "Systematic review of factors influencing the demand for medical and health insurance in Malaysia." *International Journal of Public Health Research* 10, no. 2 (2020).

[13] Sanogo, N. A., and S. Yaya. "Wealth Status, Health Insurance, and Maternal Health Care Utilization in Africa: Evidence from Gabon. BioMed Research International 2020; 2020: 1–12." https://doi.org/10.1155/2020/4036830

[14] Rawat, Seema, Aakankshu Rawat, Deepak Kumar, and A. Sai Sabitha. "Application of machine learning and data visualization techniques for decision support in the insurance sector." *International Journal of Information Management Data Insights* 1, no. 2 (2021): 100012. https://doi.org/10.1016/j.jjimei.2021.100012

[15] Chand, Satish, and Yu Zhang. "Learning from machines to close the gap between funding and expenditure in the Australian National Disability Insurance Scheme." *International Journal of Information Management Data Insights* 2, no. 1 (2022): 100077. https://doi.org/10.1016/j.jjimei.2022.100077

[16] Schaefers, Juergen, Supriyatiningsih Wenang, Andi Afdal, Ali Ghufron Mukti, Sri Sundari, and Joerg Haier. "Population-based study on coverage and healthcare processes for cancer during implementation of national healthcare insurance in Indonesia." *The Lancet Regional Health-Southeast Asia* 6 (2022). https://doi.org/10.1016/j.lansea.2022.100045

[17] Seligman, Benjamin, Shripad Tuljapurkar, and David Rehkopf. "Machine learning approaches to the social determinants of health in the health and retirement study." *SSM-population health* 4 (2018): 95-99. https://doi.org/10.1016/j.ssmph.2017.11.008

[18] Stämpfli, Dominik, Birgit A. Winkler, Simona Berardi Vilei, and Andrea M. Burden. "Assessment of minor health disorders with decision tree-based triage in community pharmacies." *Research in Social and Administrative Pharmacy* 18, no. 5 (2022): 2867-2873. https://doi.org/10.1016/j.sapharm.2021.07.003

[19] Thotad, Puneeth N., Geeta R. Bharamagoudar, and Basavaraj S. Anami. "Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 17, no. 1 (2023): 102690. https://doi.org/10.1016/j.dsx.2022.102690

[20] Trein, Philipp, Michel Fuino, and Joël Wagner. "Public opinion on health care and public health." *Preventive Medicine Reports* 23 (2021): 101460. https://doi.org/10.1016/j.pmedr.2021.101460

[21] Xie, Yuan-tao, Juan Yang, Chong-guang Jiang, Zi-yu Cai, and Joshua Adagblenya. "Incidence, dependence structure of disease, and rate making for health insurance." *Mathematical Problems in Engineering* 2018 (2018). https://doi.org/10.1155/2018/4265801