Received: 03 Jul 2024 | Revised: 10 Oct 2024 | Accepted: 07 Dec 2024 | Published: 21 Dec 2024
Volume 2(2) Pages 55-66 DOI: https://doi.org/ 10.56279/jicts.v2i2.111

JICTS
Journal of ICT Systems

Forecasting Pediatric Medical Expenses using Machine Learning: A Case Study of the Toto Afya Card in Tanzania

Fransisca Mwakinyali [1], Ruthbetha Kateule, Mahadia Tunga

[a]*Department of Computer Science and Engineering, University of Dar es Salaam, Dar es Salaam, Tanzania*

[1]Corresponding author
Email: mwakinyalif@gmail.com

Funding information

This work is self-funded.

Keywords

Artificial Intelligence
Machine Learning
National Health Insurance Fund
Medical Expenses

Abstract

The study aimed to predict medical expenses using machine learning (ML) algorithms to improve accuracy and efficiency in healthcare cost estimation. Previously, medical expenses were determined through actuarial analyses, manual assessments, or linear models based on historical data. However, such methods often fail to account for complex relationships among the numerous variables involved in healthcare costs, leading to less precise predictions. The use of ML offers a potential solution by capturing these complex interactions. This study explored four ML models, namely Linear Regression, Random Forest, XGBoost, and CatBoost, - to predict medical expenses using a dataset that included socio-demographic and healthcare-related factors. The selection of these algorithms was based on their ability to handle large datasets, non-linear relationships, and categorical features. The results show that CatBoost and XGBoost perform better than traditional methods. The study also discussed challenges posed by socio-economic conditions, healthcare infrastructure, and demographic variations in modeling.

1. Introduction

Medical expenses are one of the major recurring expenses in human life. It is well understood that one's lifestyle and various physical parameters dictate the diseases or ailments that one can have, and these ailments dictate medical expenses [1].

Medical expenditures are related to the diagnosis, treatment, mitigation, or prevention of diseases that can affect any organ or system of the body. These costs cover the payment of doctors, surgeons, dentists, and other medical professionals who provide legitimate medical services. The primary purpose of medical care cost is to treat or prevent a physical or mental illness or condition [2].

Globally, people face the challenge of high medical expenses, especially children, which can be influenced by factors like how healthcare is set up, government rules, and cultural habits [3]. In different countries, individuals and families deal with the cost of healthcare in various ways, either by paying directly or contributing to health insurance plans. The increasing use of advanced technology, higher prices for medicines, and a growing need for specialized medical services contribute to rising medical costs worldwide [3].

Around the world, children's medical expenses are a significant concern for families and communities. The cost of healthcare for children varies across countries due to differences in healthcare systems, economic conditions, and access to medical services. According to WHO [5], families often engage with the financial burden of providing healthcare for their children, covering expenses related to diagnosis, treatment, and preventive measures. In some regions, children's health may receive limited attention from healthcare services researchers, potentially impacting the overall well-being of the younger population [5].

Governments and organizations globally implement various strategies to address children's medical expenses [6]. In Tanzania, for instance, the National Health Insurance Fund (NHIF) plays a crucial role by offering the “Toto Afya” package, a health insurance scheme specifically aimed to make healthcare available to all children under the age of 18 [7]. The “Toto Afya” package is a health insurance scheme for children under 18 with an annual charge of TShs. 50,400 per child, allowing them to get medical treatment at NHIF-registered institutions around the country. The goal was to allow all Tanzanians who have not yet benefited from the fund to register their children [8].

This initiative aimed to make healthcare accessible to all children, ensuring they can receive medical treatments at registered institutions across the country [7]. The effort to alleviate children's medical expenses requires collaborative efforts, including policies that prioritize pediatric healthcare, international support, and initiatives to identify and support at-risk children in both clinical and community settings.

Predicting medical expenses for children in Tanzania presents several challenges due to a combination of factors, including socio-economic conditions, healthcare infrastructure, and demographic variations. One of the primary challenges is the diverse range of health conditions that children may face, making it difficult to accurately estimate future medical costs [8]. Children's health is influenced by various factors, such as genetics, environmental conditions, and lifestyle, which can lead to a wide spectrum of illnesses and medical needs [9].

Addressing these challenges requires a holistic approach that considers the diverse health landscape, socio-economic disparities, and the effectiveness of healthcare initiatives such as ML provides a wide range of instruments, methods, and systems [10]. Improved data collection, advanced analytics, and a comprehensive understanding of the unique healthcare dynamics in Tanzania are essential for developing accurate predictions and ensuring effective allocation of resources to support the health and well-being of children [11].

Therefore, the main objective of this study was to develop a ML based model for predicting children’s medical expenses by using insurance data, considering the unique factors influencing pediatric healthcare costs using a Toto Afya insurance scheme in Tanzania. The study utilized four common ML algorithms that have been used for predicting medical expenses, mainly: Linear Regression, CatBoost, Random Forest, and XGBoost, in predicting pediatric medical expenses. The study will contribute valuable insights for

enhancing the efficiency of health insurance companies through informed healthcare planning, decision-making, and cost management, ultimately working towards achieving more inclusive and effective universal health care coverage for children.

2. Related Work

The successful implementation of ML algorithms to predict healthcare costs in routine settings could be undermined if their predictive performance is poor or leads to overlay optimistic predictions [12]. Various elements in the design, conduct, and analysis of ML algorithm may introduce bias, including the lack of internal validation to prevent overfitting, unrepresentative sampling, or unaccounted missing data. The utility of these models may also be adversely affected by poor or inadequate reporting of the studies in the increasing body of literature through which they are disseminated to potential users including payers, health systems, and individuals. One of the studies has revealed that the methodological and reporting quality of ML-based prediction models for clinical outcomes is suboptimal [13]. However, predicting medical expenses requires different algorithms to be built so they can be used for prediction. There are four common algorithms that are mostly used by researchers for predicting medical expenses: Linear Regression, Random Forest, XGBoost, and CatBoost.

Taloba et al. [15] Focused on predicting adult's medical expenses and estimating hospitalization. Obesity was predicted to raise healthcare expenses. The methodology of such research was divided into two types depending on the number of variables in the model: simple linear regression and multiple linear regression. The linear regression algorithm was used to estimate healthcare costs of patients, such as obesity (body mass index, BMI) using certain devices, such as smartphones and smart devices. The study was based only on the prediction of medical expenses for people with obesity, and used one algorithm for the prediction and its accuracy was 70.1%.

Raita et al. [16] developed four ML models: Lasso regression, random forest, gradient-boosted decision tree, and deep neural network. As the reference model, this study constructed a logistic regression model using the five-level ESI data. The clinical outcomes were critical care (admission to the intensive care unit or in-hospital death) and hospitalization (direct hospital admission or transfer). Based on this study, data were collected from the National Hospital and Ambulatory Medical Care Survey (NHAMCS) ED data from 2007 through 2015. The study identified all adult patients only (aged above18 years), but the data did not collect some helpful clinical variables (e.g., chronic medications, socioeconomic status, health behaviors), and the quality of the data was low, which led to 69.8% accuracy.

Huang et al. [17] focused on employing feature selection to determine the critical characteristics that influence postoperative spending. Also, ML algorithms to construct an acceptable medical expense predictive model for coronary artery bypass grafting (CABG) patients. This study used five types of ML in the training set to train by selecting the relevant features for medical expense prediction, including LR, classification and regression tree (CART), support vector regression (SVR), multivariate adaptive regression splines (MARS), and XGBoost (extreme gradient boosting). Limitation was only applicable for people above 18 years who undergo CABG surgery, and their model could find that the corresponding operation variables to predict one-year medical expenditure after CABG surgery.

Muremyi et al. [17] used ML approaches and compare the results using four ML approaches, such as random forest, decision tree models, gradient boosting, and regression tree models. The

data to use for analysis was collected from the National Institute of Statistics (NISR), which is the Integrated Living Conditions Survey 2016–2018 (EICV5). Gradient boosting was selected as the best model because the train accuracy was 78% and the test accuracy was 85%. Information was collected at the household and individual levels above 18 years based on diseases, such as diabetes and high blood pressure. The findings demonstrated that medical expenses are significantly related to age and healthcare expenditures. ML models can help accurately forecast expenditures. These results could advance the field toward precise preventive care to lower overall healthcare costs and deliver care more efficiently. The limitation is that the study is based only on people with diabetes and high blood pressure.

Sohn et al. [18] developed regression models to predict healthcare expenditures, and binary classification models were developed to predict whether a participant's healthcare expenditure was in the top 50%. Both regression and classification models were developed in four versions: a baseline model (T for "tabular data") that relies only on patient sex, age, and ZIP code median income as indicators; healthcare spending data was obtained from the cost accounting unit of the institution's hospital financial department; the study focused only on adult chest radiographs data.

Kodiyan and Francis [19] focused on finding the correlation of medical expenses with each of the attributes and using these attributes to predict charges. The forecasting medical costs dataset was obtained from people's Kaggle accounts and consists of information and annual insurance premiums provided to them. The methodology used was multiple regression and then ANOVA to compare different models and find the best-fit model. The regression could be predicted with more than 75% accuracy charges.

Utilizing the medical information and costs dataset from Kaggle, performed a predicate analysis on the medical health insurance costs of adult based on gender, age, smoking habits, BMI, number of children, and region. The study applied multiple linear regression, support vector regression, decision tree regression, and random forest regression models to the dataset using ML techniques. The study's findings showed that random forest regression outperforms the other three algorithms. Age was additional factor that had a significant impact on medical insurance costs [20].

Prior to the application of ML, medical expenses were typically predicted using simpler, linear approaches, such as regression analysis, or through actuarial tables, which often ignored the influence of non-linear relationships and interaction terms between variables. The current status of expense prediction mostly relies on statistical methods or expert-driven estimations. The shift to ML enables the integration of diverse and large datasets, capturing more complex relationships and thus improving predictive accuracy, particularly in contexts where socio-economic conditions, healthcare infrastructure, and diverse health conditions differ widely across regions.

3. Methodology

This study was categorized as applied experimental research, which aims to apply or extend theories to solve practical real-world problems with a desired practical outcome. The study employed ML approaches to predict children's medical expenses, with a focus on analyzing insurance data, specifically the "Toto Afya" package.

3.1 Data Collection

The study utilized data sourced from the National Health Insurance Fund (NHIF), focusing

on the "Toto Afya" package for the year 2022. The study delves into the spending patterns of children enrolled with the Toto Afya card, examining the alignment between their medical expenses and the corresponding charges they incurred. The dataset, representing a sample, was derived from the broader population of children in Tanzania, ensuring representation from diverse regions. Notably, the inclusion of participants from high-population regions like Dar es Salaam, Mwanza, Mbeya, and Arusha enhances the study's geographical diversity and relevance.

3.2 Data Pre-processing

Data preprocessing involved handling of missing values and data splitting. Exploratory Data Analysis (EDA) was used to prepare the data for further analysis and model development; Python libraries such as Pandas and NumPy were utilized.

3.2.1 Data Cleaning and Preparation

The data was cleaned by using Python libraries, which were Pandas and Matplotlib. Removing missing values and duplicate values was an important part of data pre-processing. Missing values can be a problem for many types of analysis and ML model development, and can lead to biased or inaccurate results. Duplicate values can also be a problem for data analysis, as they can skew summary statistics and lead to overestimation of the size of the dataset.

3.2.2 Checking for Multicollinearity

In this study, correlation analysis was used to check multicollinearity by quantifying the degree of association between independent variables. High correlations indicate potential multicollinearity, a condition where predictor variables are highly interrelated. Identifying and addressing multicollinearity through correlation analysis was crucial during data preprocessing to ensure stability and interpretability of ML models, preventing issues such as inflated standard errors and unreliable coefficient estimates. Correlation analysis was used to measure the strength and direction of the relationship between variables. Computed the correlation coefficients and created heat maps to visualize the correlations. Kernel Density Estimation plots were used to visualize the variables' probability density distributions and identify the distributions' modes and shapes.

3.2.3 Exploratory Data Analysis

The study employed the Exploratory Data Analysis (EDA) techniques: descriptive statistics to summarize key features in the dataset, providing insights into central tendencies and variability; multivariate analysis to explore relationships among multiple variables simultaneously, aiding in understanding complex interactions; outlier detection to identify and address extreme values that could skew analyses, ensuring data integrity and robustness in subsequent modeling. To prepare the data for further analysis and model development, Python libraries such as Pandas and NumPy were utilized.

3.2.4 Feature Engineering

Feature engineering is crucial in ML for improving model performance and interpretability. Techniques which were used in this study include creating new features, handling missing data through imputation, checking collinearity, and transforming variables to better align with model assumptions. Therefore, the study uses Python libraries, such as Pandas and NumPy to create new features, check multicollinearity, and handle missing data.

3.3 Model Development

The study employed four different machine learning (ML) algorithms to identify the best-performing models. These algorithms were chosen based on their widespread usage in predicting medical expenses [21].

The algorithms chosen for this study were selected due to their varied capabilities in handling non-linear relationships, feature importance, and categorical data. Linear regression served as the baseline, representing traditional models. Random forest was selected due to its ability to capture non-linear relationships and handle high-dimensional data, while XGBoost and CatBoost were chosen for their state-of-the-art performance in predictive tasks, particularly in cases with categorical variables and imbalanced datasets.

The performance differences were analyzed, with CatBoost and XGBoost outperforming due to their ability to better handle categorical data and avoid overfitting. (Figure 1) presents the workflow for developing the ML models. The workflow included training the ML algorithms using the prepared data, which were split using the train-test split principle. The process comprised partitioning the data into two distinct subsets, namely the training and testing sets. The model was fitted to 80% of the available data to adjust its parameters, while the remaining 20% was reserved for evaluating its efficacy in predicting outcomes on novel and previously unseen data.

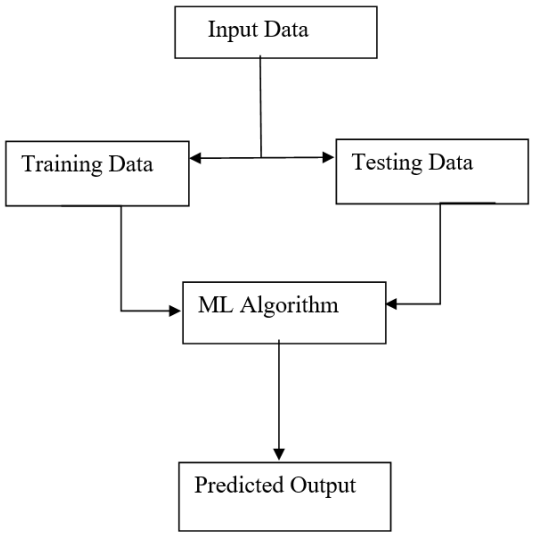

Figure 1. ML model training workflow.

4. Results

The results delve into the analysis and interpretation of a study aimed at predicting medical expenses for children. The data pre-processing steps were undertaken to prepare the dataset for analysis. Key findings from the dataset are highlighted, including insights into age distribution, regional contributions, ownership types of medical facilities, and correlations between visit frequency and medical expenses.

The analysis reveals that most children receiving medical services were males (Figure 2).

Figure 2. Distribution of male and female.

Region: The most region that made a huge contribution was Kinondoni in the number of visitors. Out of 23 regions it is only three Kinondoni, Temeke and Ilala demonstrated a substantial difference in visitor as shown in (Figure 3).

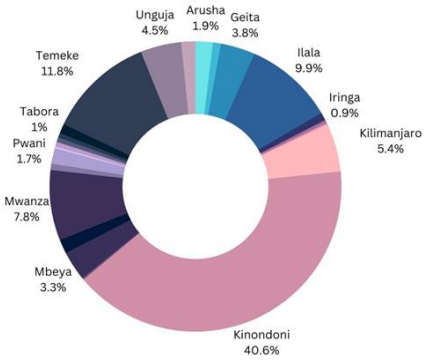

Figure 3. Regions distribution.

Ownership: The most was a private hospital where the ownership consists of the hospital, clinic, or dispensary where children received medical services. The ownership can be public, private, or faith-based as shown in (Figure 4).

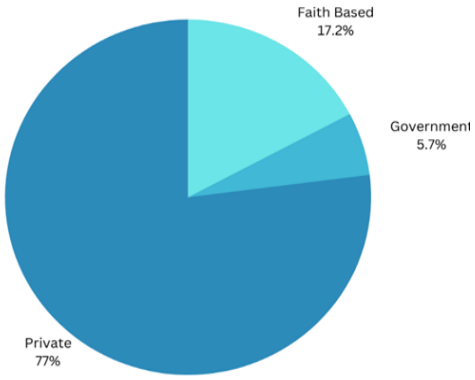

Figure 4. Distribution of ownership.

A positive correlation between the number of visits and medical expenses is noted, suggesting that frequent visits lead to higher expenses. Moreover, regional differences impact both visitation patterns and the average amount spent on medical services. For instance, patients in Kinondoni visited more frequently compared to those in regions like Katavi, possibly due to varying healthcare accessibility and patient income levels.

There was no substantial difference in expenses between male and female patients, indicating that gender does not significantly influence healthcare costs for children. However, factors like visit frequency, age group, and regional disparities play a more significant role in determining medical expenses.

CatBoost emerges as the top performer with an accuracy score of 82.1%, followed by XGBoost, Random Forest, and Linear Regression models. The evaluation includes metrics like F1 score, highlighting the superior performance of CatBoost and XGBoost over other models (Figure 5).

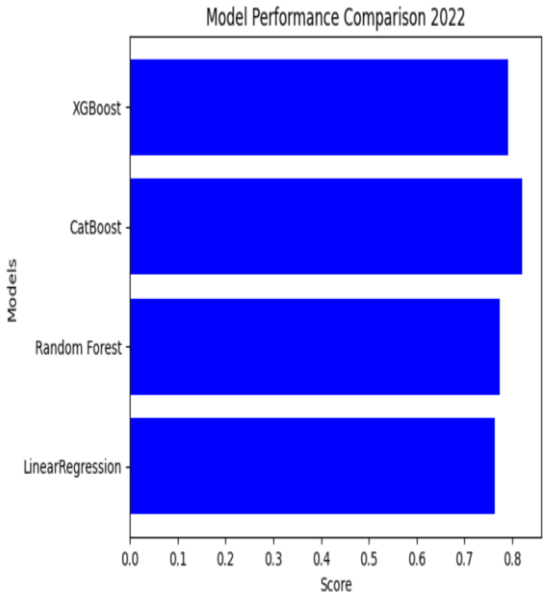

Figure 5. Performance measure of amount paid.

The model accuracy on the test set emphasizes the effectiveness of CatBoost and XGBoost in predicting medical expenses. After eliminating unimportant variables, CatBoost stands out as the most accurate method for predicting insurance costs for individuals. The comparison of actual against predicted values showcases the robustness of these models in capturing the variation in medical expenses (Table 1).

XGBoost demonstrates the highest ability to explain variation (R2), followed closely by CatBoost. The Root Mean Squared Error (RMSE) estimate further validates the superiority of XGBoost in predicting medical expenses accurately.

Table 1. Example of actual values versus predicted values.

	Actual Values	Predicted Values
412	12950	2146.219365
96	7780	3781.892899
478	54000	49635.412489
426	14950	33291.668318
564	34150	36853.360421

The implications of these results are significant for the healthcare industry and insurance providers. The high accuracy of the predictive models indicates the potential for leveraging ML algorithms to forecast children's medical expenses. This can enable healthcare providers and insurance companies to make informed decisions, plan interventions, and allocate resources more effectively. Additionally, the accurate predictions can aid in developing tailored healthcare packages and insurance plans, ultimately leading to improved healthcare transparency and cost management.

The superior performance of CatBoost and XGBoost models can be attributed to their inherent characteristics and the nature of the dataset. CatBoost, known for its robust handling of categorical variables and robustness to overfitting, may have excelled due to the presence of categorical features in the dataset. XGBoost, with its ability to handle missing data and outliers effectively, might have performed well due to the dataset's characteristics. The ensemble nature of these models, combining multiple weak learners to form a strong learner, also contributes to their high predictive power.

The high performance of CatBoost and XGBoost models can be scientifically explained by their underlying algorithms. CatBoost's ability to handle categorical features and its robustness to overfitting make it well-suited for datasets with diverse categorical variables and complex interactions. XGBoost gradient boosting framework, with its capacity to handle missing data and outliers, allows it to effectively capture the complex relationships within the dataset, leading to accurate predictions.

5. Discussion

The study's findings provide significant insights into the factors influencing medical expenses and the effectiveness of various ML models in predicting these costs. The dataset revealed important demographic and regional insights that impact healthcare costs. Notably, male children received more medical services than females, and younger children, particularly those aged 5 years, were the most frequent visitors to healthcare centers. This age group had higher healthcare interactions compared to adolescents aged 13 to 17 years. Furthermore, regional variations showed that Kinondoni, Temeke, and Ilala had significantly higher numbers of healthcare visits, indicating disparities in healthcare service utilization.

Numerous approaches have been implemented to constrain the growth of healthcare spending, such as capitated payments and value-based insurance designs. However, these methods often depend on accurate predictions of future health spending, which is crucial for efficient resource allocation. Historically, regression-based methodologies have been the most popular for forecasting health spending. These parametric techniques, while useful, often struggle with sparse or missing data due to their significant assumptions about data generation mechanisms [22]. Recently, advancements in computing have led to an increased use of ML approaches in predicting health spending, including classifying patients based on predicted spending amounts or changes over time. However, the successful implementation of ML algorithms in routine settings is contingent on their predictive performance, which can be compromised by biases, poor internal validation, unrepresentative sampling, or inadequate reporting [23].

Among the evaluated models, CatBoost demonstrated the highest accuracy (82.1%) on the training set, outperforming XGBoost (79.2%), Random Forest (77.3%), and Linear Regression (76.4%). This indicates CatBoost's robustness in handling the complexities of the dataset. Both CatBoost and XGBoost also had higher F1 scores compared to the other models, further validating their effectiveness in predictive performance.

On the testing set, CatBoost remained the most accurate model after eliminating less significant variables. XGBoost showed the highest R^2 value, indicating its strong explanatory power for variations in medical expenses. The superior RMSE estimate for XGBoost further validated its effectiveness in minimizing prediction errors. These results highlight that both CatBoost and XGBoost are highly effective ML algorithms for predicting medical expenses.

The high performance of CatBoost and XGBoost models in this study suggests that these algorithms are well-suited for predicting healthcare expenses. Stakeholders, including healthcare providers, policymakers, and insurance companies, can leverage these models to forecast medical costs accurately and plan interventions effectively. By utilizing the insights from these predictive models, stakeholders can make more informed decisions regarding resource allocation, preventive care strategies, and financial planning, ultimately aiming to optimize healthcare delivery and control costs. Future research should continue to refine these models and explore additional variables to further improve the accuracy and applicability of healthcare cost predictions.

The study acknowledges that socio-economic conditions, health care infrastructure, and demographic variations significantly impact medical expenses. These factors were incorporated through variables related to income, healthcare access, age, and region. By including these variables in the model, we aimed to account for the differences in healthcare utilization patterns across different demographic and socio-economic groups. The ML models were specifically tailored to learn from such variations, and the use of advanced algorithms like CatBoost and XGBoost further helped to manage the categorical data representing these factors, allowing for a more accurate prediction.

6. Conclusion

The study contributes valuable insights into healthcare cost prediction, highlighting the importance of data quality, variable selection, and regional analysis in accurate modeling. Also, the study underscores the potential of advanced ML algorithms in forecasting healthcare costs and the critical role of demographic and regional factors in influencing medical expenses. By harnessing these insights, stakeholders can make strategic decisions to improve healthcare outcomes and sustainability. For stakeholders, including healthcare providers, policymakers, and insurance companies, the insights gained from these predictive models can facilitate more informed decision-making. By leveraging these models, stakeholders can optimize resource allocation, develop preventive care strategies, and improve financial planning, ultimately aiming to enhance healthcare delivery and control costs.

Future research should focus on refining these ML models and incorporating additional variables to further improve their accuracy and applicability in various healthcare contexts. Continuous improvement in predictive modeling will enable better management of healthcare expenses and contribute to more efficient and equitable healthcare systems.

ACKNOWLEDGEMENT

Authors acknowledge the National Health Insurance Fund of Tanzania for providing the data and supplementary research resources utilized in this study.

CONTRIBUTIONS OF CO-AUTHORS

Fransisca Mwakinyali		Conceived the idea, conducted data collection, performed data analysis, data interpretation and wrote the paper
Ruthbetha Kateule	[ORCID: 0000-0002-0413-3981]	Conceived the idea, performed data interpretation, wrote and reviewed the paper
Mahadia Tunga	[ORCID: 0000-0003-2496-3373]	Provided technical info on methods and materials

REFERENCES

[1] IRS, *Medical and Dental Expenses 2022*. Available: https://www.irs.gov/pub/irs-prior/p502--2022.pdf.

[2] United Nations, *General Assembly Resolution 70/1- Transforming our world: the 2030 Agenda for Sustainable Development*. 2015. Available:https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf.

[3] National Academies of Sciences, Engineering, and Medicine, *Approaches to Universal Health Coverage and Occupational Health and Safety for the Informal Workforce in Developing Countries*, Workshop Summary. Washington, DC: The National Academies Press, 2016. Available: https://doi.org/10.17226/21747.

[4] WHO, *Towards a Healthier Nation: WHO Investment Case for 2020-2021*, 2020. Available: https://www.afro.who.int/sites/default/files/2021-03/WHO%20Investment%20Case_November%202020.pdf.

[5] B. Lee, K. Tarimo, and A. Dutta, *Tanzania's Improved Community Health Fund an Analysis of Scale-Up Plans and Design*. HP Policy Brief, 2018.

[6] PATH, *Child Health in Tanzania: Identifying policy pathways to help prevent child deaths from pneumonia and diarrhea*. 2014. Available: https://media.path.org/documents/APP_tanzania_pneumo_dd_brief.pdf.

[7] NHIF. *The National Health Insurance Fund Act*, Revised Edition of 2015. Available: https://www.kazi.go.tz/uploads/documents/en-1600098938-NHIF-ACT-1.pdf.

[8] R. F. Msacky, *Quality of health service in the local government authorities in Tanzania: a perspective of the healthcare seekers from Dodoma City and Bahi District councils*. BMC Health Serv Res. 24, Vol 1, pp 81, 2024.

[9] J. F. Outreville, *Theory and Practice of Insurance*. Theory Pract. Insur., no. June 2016, 1998. Available: https://doi.org/10.1007/978-1-4615-6187-3.

[10] T. Pfutzenreuter and E. Pinheiro de Lima, *Machine Learning in Healthcare Management for Medical Insurance Cost Prediction*. Available: 10.14488/ENEGEP2021_TI_ST_354_1820_42095.

[11] A. O. Abiodun. *The Impact of Advertising in Improving Sales Volume of a New Product: A case study of Starcomms Plc, Nigeria*. HAMK University of Applied Sciences, 2019. Available: https://www.theseus.fi/bitstream/handle/10024/34928/Adekoya_Olusola.pdf.pdf.

[12] Prayitno, Shyu CR, Putra KT, Chen HC, Tsai YY, Tozammel Hossain KSM, et al., *A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications*. Applied Sciences (Switzerland), vol 11, 2021.

[13] A. W. Huang, M. Haslberger, N. Coulibaly, O. Galárraga, A. Oganisian, L. Belbasis, et al., *Multivariable prediction models for health care spending using machine learning: a protocol of a systematic review*. Diagn Progn Res. Vol 6, 2022.

[14] A. I. Taloba, R. M. A. El-aziz, and H. M. Alshanbari, *Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning* 2022;2022.

[15] Y. Raita, T. Goto, M. K. Faridi, D. F. M. Brown, C. A. Camargo, and K. Hasegawa, *Emergency department triage prediction of clinical outcomes using machine learning models*. Crit Care, vol 23, pp 1-13, 2019.

[16] Y. Huang, S. Li, M. Chen, and T. Lee, *The Prediction Model of Medical Expenditure Appling Machine Learning Algorithm in CABG Patients*, Healthcare (Basel), vol 9, issue 6, pp 710, 2021.

[17] R. Muremyi, N. Francois, K. Ignace, N. Joseph, and D. Haughton, *Comparison of Machine Learning Algorithms for Predicting the Out-of-Pocket Medical Expenditures in Rwanda*. J Health Med Res, vol 1, pp 32 - 41, 2019.

[18] J. H. Sohn, Y. Chen, D. Lituiev, J. Yang, K. Ordovas, D. Hadley, et al. *Prediction of future healthcare expenses of patients from chest radiographs using deep learning: a pilot study*. Sci Rep. vol 12, pp 1 – 9, 2022.

[19] A. A. Kodiyan, and K. Francis, *Linear regression model for predicting medical expenses based on insurance data 2019.* Available: https://doi.org/10.13140/RG.2.2.32478.38722.

[20] A. Lakshmanarao, C. S. Koppireddy, and G. V. Kumsar, *Prediction of medical costs using regression algorithms.* Journal of Information and Computational Science, vol 10, pp 5, 2020.

[21] S. Choi, and J. Blackburn, *Patterns and Factors Associated with Medical Expenses and Health Insurance Premium Payments*. Journal of Financial Counseling and Planning, vol 29, pp 6-18, 2018.

[22] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, *Machine learning approaches for predicting high-cost high need patient expenditures in health care.* BioMed Eng OnLine 17, vol 1, pp 131, 2018.

[23] U. Orji and E. Ukwandu, *Machine learning for an explainable cost prediction of medical insurance.* Machine Learning with Applications, vol 15, 2024.