Medical Insurance Price Prediction Using Ensemble Machine Learning Algorithms

Prof. Gajanan Gambhire
Mechanical Engineering
Vishwakarma Institute of Technology
Pune, India
gajanan.gambhire@vit.edu

Samruddhi Khillari
Computer Science and Engineering (AI&ML)
Vishwakarma Institute of Technology
Pune, India
samruddhi.khillari24@vit.edu

Aarya Sadawrate
Computer Science and Engineering (AI&ML)
Vishwakarma Institute of Technology
Pune, India
aarya.sadawrate24@vit.edu

Rutuja Varpe
Computer Science and Engineering (AI&ML)
Vishwakarma Institute of Technology
Pune, India
rutuja.varpe24@vit.edu

Aman Shaikh
Computer Science and Engineering (AI&ML)
Vishwakarma Institute of Technology
Pune, India
aman.shaikh24@vit.edu

Rajvardhan Desai
Computer Science and Engineering (AI&ML)
Vishwakarma Institute of Technology
Pune, India
rajvardhan.desai24@vit.edu

***Abstract*— [TODO]**

Keywords*— *insurance

I. Introduction

These days, healthcare and medical treatments are incredibly important. The cost for healthcare is also increasing globally. Since the COVID-19 pandemic, people everywhere are more aware of health and the need to be prepared for unexpected medical costs. Because of this, many people want health insurance for their families. But with so many options, it can be hard to know which one to choose. That's why we're working on a project to predict health insurance costs. This will help people understand what they might have to pay and choose the best insurance for their budget.

Medical insurance is important for everyone regardless of their life stage. Adults require chronic disease coverage as in 2016 estimation 422 million people have diabetes. In the case of middle-aged individuals, they are at a higher cancer risk. WHO reported in 2021 that 2.5 million deaths occur annually due to preventable diseases[1]. These include facts that (0-12 years) need coverage for vaccinations and emergency care, Adolescents (13-25 years) face rising mental health issues. With its reporting 10 million cases globally in 2021. Amongst the seniors (60+ years), most need long-term care. As per a 2022 UNICEF report [2], 1 in 7 suffer from disorders. An OECD report [3] found that over 50% of seniors require assisted healthcare. This highlights the importance of the right medical insurance regardless of the stage of life an individual is in.

This study aims to develop a predictive model for estimating health insurance costs, helping people to make right financial and health related decisions. We're using publicly available information from Kaggle[ ]. This information includes attributes like age, weight, whether someone smokes, how many children they have, and if they're a man or woman. These things have a big impact on how much insurance costs, as stated in previous research that we studied. We're using machine learning algorithms, to look at this data and build a model that can predict costs. By teaching the model with past insurance information, we aim to accurately guess what future insurance will cost.

The organization of the paper is as follows, Section 2 presents a comprehensive review of related work and existing literature. Section 3 details the dataset utilized in the study. In Section 4, the process of model development and its evaluation is explained. Section 5 provides an in-depth discussion of the results and their implications. Section 6 outlines the limitations of the study and explores possible directions for future research. Finally, Section 7 concludes the paper by summarizing the key findings, along with a reiteration of the limitations and future scope.

II. Literature Survey

To study existing literature/solutions, we have searched "medical insurance cost prediction," "health insurance premium prediction," "machine learning in insurance," "Regression models for insurance costs," "Predictive modeling in healthcare insurance," and "Healthcare cost estimation" keywords or syntax on Google Scholar, IEEE, and ResearchGate, and we identified several results, out of which we have chosen 20 research papers that we have open access to. So we have summarized insights from the selected studies based on important and relevant parameters like objectives of study, input parameters, ML algorithms, data type, most influential parameters and limitations of these studies. Selected papers are published in the time span of 2020 to 2024.

In order to understand objectives of published articles, we have categorized them into five different categories. The first category of objective focuses on comparison of performances of multiple algorithms. Ensemble and kernel-based methods such as XGBoost, Random Forest, and Support Vector Machine were evaluated in [3] and [4]. Linear Regression and Support Vector Machine were compared in both papers [4] and [15], providing insight into the performance difference between linear and non-linear approaches. The combination of XGBoost and Random Forest appeared across multiple studies [3], [4] and [17]. The comparison of boosting algorithms and instance-based learning such as XGBoost and k-Nearest Neighbors are compared in [3] and [20]. The second category focuses on methodology enhancement to existing methodologies for more accurate predictions. Through attribute combination analysis, efficiency-based model selection, and domain-specific data cleaning techniques that led to achieving 99.5% accuracy with their Gradient Boosting Regression model in [1]. Paper [14] demonstrated faster data processing capabilities for handling large medical insurance datasets through a big data approach

using Apache Spark with gradient-boosted tree regression which performs well with 0.9067 as R2 value.

The reviewed papers employ regression techniques to predict health insurance cost. Total 19 distinct algorithms have been used among them Random Forest Regressor Performed best in 6 studies as accuracy of 84% in [6] and [13] and R2 value as 0.9533 in [12], 0.887 in [16] and 0.85388 in [4] and [10]. XGBoost out performed in 3 studies with R2 value as 0.8681 in [11],[20] and 0.8647 in [17]. Gradient Boosting variants were appeared as best in 6 as studies accuracy of 99.5% in [1], 86% in [2], 86.82% in [3], 86.86% in [9], and R2 value as 0.9067 in [14] and 0.89 in [15].

The foundation of any prediction model rests on its input parameters, All 20 studies utilized age, sex/gender, BMI (Body Mass Index), Smoker status as an input parameter. Apart from this, several studies include additional parameters to enhance model accuracy such as pre-existing conditions (diabetes, blood pressure) [17], lifestyle factors (alcohol consumption, exercise habits) [12], medical history including previous surgeries [17] and hospitalization [12], and health behaviours such as regular check-ups and daily steps [12]. Other studies frequently include parameters like number of children and region, reflecting demographic and geographic influences on insurance costs. These additional inputs highlight the diverse approaches to improving predictive performance across the studies.

There are certain input parameters that have a strong influence on a person's insurance cost. smoking status identified smoking status in 12 studies as the most influential parameter followed by Age and BMI ranked highly influential in 11 studies. Additionally, the number of children and region (northeast) as influential in [5] while study [12] points weight as a key determinant. These parameters highlight the critical factors driving accurate predictions across the studies.

The reviewed studies demonstrate significant progress in predicting medical insurance costs However, recurring limitations highlight gaps in current research. Small datasets, as noted in Papers [1], [13], and [20] relied on small datasets with 1338 rows affecting real world applicability. Limited algorithm diversity [7] which relies on a single algorithm. Computational complexity, as identified in Papers [3] and [5], affects real-time deployment, while lack of interpretability, raised in [4], [8] and [17], poses challenges for stakeholder trust. These constraints underscore the need for future studies to leverage larger datasets, explore advanced models like deep learning [11], [20], prioritize interpretability [4], and ensure real-world validation [17]. Addressing these limitations will enhance the accuracy, fairness, and practical utility of predictive models in the insurance domain.

III. Dataset

[TODO]

IV. Methodology

[TODO]

V. Results

[TODO]

VI. Future Scope & Limitation

[TODO]

VII. Conclusion

[TODO]

References

[1] World Health Organization, "The top 10 causes of death," WHO.int, Dec. 9, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] United Nations Children's Fund, "Mental health," UNICEF Data, https://data.unicef.org/topic/child-health/mental-health/

[3] OECD, *Is Care Affordable for Older People?*, OECD Health Policy Studies, OECD Publishing, Paris, 2024. [Online]. Available: https://doi.org/10.1787/450ea778-en

[4] N. Bhardwaj and R. Anand, "Health insurance amount prediction," *Int. J. Eng. Res. Technol.*, 2020.

[5] A. U. Hassan, J. Iqbal, S. Hussain, M. A. A. Mosleh, and S. S. Ullah, "A computational intelligence approach for predicting medical insurance cost," *Math. Probl. Eng.*, vol. 2021, pp. 1–13, 2021.

[6] M. Hanafy and O. M. A. Mahmoud, "Predict health insurance cost by using machine learning and DNN regression models," *Int. J. Innov. Technol. Explor. Eng.*, 2021.

[7] V. Ramachandran, A. R. Kavitha, and P. R., "An accurate prediction of medical insurance cost using forest regression algorithms," in *Proc. 2023 Int. Conf. Data Sci., Agents, Artif. Intell.*, 2023.

[8] N. Shakhovska, N. Melnykova, V. Chopiyak, and M. Gregus, "An ensemble methods for medical insurance costs prediction task," *CMC-Comput. Mater. Continua*, 2021.

[9] D. Ramya, K. M. S., and J. Deepa, "Health insurance cost prediction using machine learning algorithms," in *Proc. 2022 Int. Conf. Edge Comput. Appl. (ICECAA)*, 2022.

[10] K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar, and R. K. Bhatia, "Health insurance cost prediction using machine learning," in *Proc. 2022 3rd Int. Conf. Emerg. Technol. (INCET)*, 2022.

[11] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine learning-based regression framework to predict health insurance premiums," *Int. J. Environ. Res. Public Health*, 2022.

[12] M. Kulkarni, "Medical insurance cost prediction using machine learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, 2022.

[13] A. Bharti and L. Malik, "Regression analysis and prediction of medical insurance cost," *Int. J. Creative Res. Thoughts*, 2022.

[14] S. Hossen, "Medical insurance cost prediction using machine learning," *ResearchGate*, 2023. [Online]. Available: https://www.researchgate.net. [Accessed: Apr. 13, 2025].

[15] V. Vijayalakshmi, A. Selvakumar, and K. Panimalar, "Implementation of medical insurance price prediction system using regression algorithms," in *Proc. 2022 3rd Int. Conf. Emerg. Technol. (INCET)*, 2023.

[16] T. Thejeshwar, T. S. Harsha, V. V. Krishna, and R. Kaladevi, "Medical insurance cost analysis and prediction using machine learning," in *Proc. 2023 Int. Conf. Innov. Data Commun. Technol. Appl. (ICIDCA)*, 2023.

[17] S. Albalawi, L. Alshahrani, and N. Albalawi, "Prediction of healthcare insurance costs," *Comput. Inform.*, vol. 3, no. 1, 2023.

[18] J. A. S. Cenita, P. R. F. Asuncion, and J. M, "Performance evaluation of regression models in predicting the cost of medical insurance," *Int. J. Comput. Sci. Res.*, 2023.

[19] K. L. Narayana, "Medical insurance premium prediction using regression models," *Int. J. Res. Technol. Innov.*, 2023.

[20] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," *Elsevier*, 2024.

[21] A. Reddy and L. Madhuri, "Medical health insurance price prediction," *Int. J. Novel Res. Dev.*, 2024.

[22] S. U. S and A. Mathew, "Medical insurance cost prediction," *Int. J. Data Commun. Netw.*, 2024. [Online]. Available: https://www.ijdcn.latticescipub.com/portfolio-item/D503704040624/. [Accessed: Apr. 13, 2025].

[23] G. K. Patra, C. Kuraku, S. Konkimalla, V. N. Boddapati, M. Sarisa, and M. S. Reddy, “An analysis and prediction of health insurance costs using machine learning-based regressor techniques,” *J. Data Anal. Inf. Process.*, 2024.