A Comprehensive Study on the Premium estimation of the Health Insurance Sector

Name: Mrs. Kalyani Gorti

Assistant Professor, Department of Commerce, Bhavan's Vivekananda College

Email ID: kalyanigorti@gmail.com

Name: Srihitha Patibanda

Student, B.com(Hons)Business Analytics , Bhavan's Vivekananda College

Email ID – srihithapatibanda@gmail.com

Name : K.V. Kanchan

Student, B.com(Hons)Business Analytics , Bhavan's Vivekananda College

Email ID: kanchankandadai@gmail.com

Name: Ratan Praneeth

Graduate, B.com(Hons)Business Analytics , Bhavan's Vivekananda College

A Comprehensive Study on the Premium estimation of the Health Insurance Sector

Abstract

Health insurance plays a crucial role in providing financial security against medical expenses, ensuring accessibility to healthcare services. However, premium estimation remains a complex process influenced by multiple factors such as age, medical history, and lifestyle choices. This study explores key determinants of health insurance premiums and employs machine learning models to enhance prediction accuracy.

The study employs secondary data and uses a range of predictive models, such as Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors, to efficiently estimate insurance premiums. Exploratory Data Analysis (EDA) and statistical methods are used to detect meaningful correlations and enhance the transparency of premium calculation. The results show that age, chronic conditions, previous surgeries, and hereditary diseases are significant drivers of premium charges. Among the models tested, Random Forest Regression demonstrated the highest accuracy in premium prediction.

By integrating machine learning into premium estimation, this study aims to improve transparency, optimize pricing structures, and empower consumers with better financial planning tools. Future research can further refine predictive models by incorporating real-time claim data and additional health-related variables.

Keywords: *Health Insurance, Premium Estimation, Machine Learning, Risk Assessment, Predictive Modeling.*

Introduction

In today's unpredictable world, health insurance acts as a vital financial shield, protecting individuals and families from the high costs of medical care. It covers expenses related to hospital stays, surgeries, and routine treatments, ensuring access to quality healthcare without imposing excessive financial strain. By paying a predetermined premium, policyholders receive varying levels of coverage, depending on the terms of their chosen policy. This makes health insurance an essential component of financial planning, offering security against unexpected medical expenses. Beyond financial protection, health insurance enhances accessibility to medical services by encouraging preventive care and early treatment. It enables individuals to seek timely medical attention, reducing long-term healthcare costs and improving overall public health. Policies can be obtained through employer-sponsored plans, government programs, or private insurers, each providing different coverage options suited to diverse needs. With the increasing prevalence of chronic diseases and rising healthcare costs, understanding premium estimation has become crucial for both insurers and policyholders. Accurate premium estimation ensures fair pricing and affordability while maintaining the financial stability of insurance providers. Various factors, including age, medical history, and lifestyle choices, influence premium costs. This study aims to explore the key determinants of health insurance premiums and evaluate predictive models that can improve estimation accuracy. By leveraging data-driven approaches, this research seeks to enhance transparency, helping individuals make informed decisions about their healthcare coverage.

Need for the Study

Insurance premium calculations have traditionally lacked transparency, relying on agents and complex pricing models that consumers struggle to understand. As healthcare costs continue to rise, individuals need more clarity in determining fair and affordable premiums. The use of machine learning and statistical analysis offers an opportunity to improve accuracy and consumer control in premium estimation. This study aims to develop a predictive model based on health metrics and medical history to provide better premium estimations. By optimizing pricing structures and reducing reliance on traditional channels, the model promotes fair, personalized insurance rates. Increased transparency will empower individuals to make informed decisions, fostering a more consumer-centric and equitable insurance sector.

Review of Literature

Iqbal J., Hussain S(2022) presented a computational intelligence technique for estimating healthcare insurance expenditures using a set of machine learning techniques. the potential outcome of using predictive

algorithms to calculate insurance prices. In the second stage, the authors looked at how the connection between the company and the insured was altered when the customer realised that the firm had a lot of data about her actual behaviour that was constantly updated. **Kumar Sharma and Sharma (2022)** aimed to develop mathematical models for predicting future premiums and validating the findings using regression models. To anticipate policyholders' choice to lapse life insurance contracts, we employed the random forest approach. Even when factoring in feature interactions, the technique beats the logistic model. Azzone et al. studied how the model works; we employed global and local classification tools. The findings suggest that existing models, such as the logistic regression model, are unable to account for the variety of financial decisions.**Hanafy M and Mahmoud O.M.A(2021)** on Predicting Health Insurance Cost by Using Machine Learning and DNN Regression Models tells us that several variables impact the cost of insurance. These elements have an impact on the development of insurance plans. Machine learning (ML) can help the insurance industry enhance the efficiency of policy wording. Forecasting insurance costs based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual. Machine learning can significantly minimize these individual efforts in policymaking, as ML models can do cost calculation in a short time, while a human being would be taking a long time to perform the same task. **Nidhi Bhardwaj and Rishabh Anand (2021)** used health data to predict insurance premiums through regression models, finding that multiple linear regression and gradient boosting outperformed other methods, with gradient boosting being more efficient. Their study emphasized the role of predictive analytics in life insurance risk assessment, utilizing a large anonymized dataset and feature selection techniques to improve accuracy. Accurate premium estimation is essential for financial planning, as errors can impact insurers' stability. Goundar et al. further demonstrated the effectiveness of artificial neural networks (ANNs) in forecasting annual medical claims using feedforward and recurrent neural network. **Dr Eline van den Broek-Altenburg(2019)** on using social media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrolment Season suggests that other, non-financial factors, might be important in the choice of health insurance plan, such as the sentiments that consumers have. In some sense, "fear" is simply "risk-averse" in the vernacular. In another sense, however, this study provides specificity about the nature of the risk aversion and suggests that consumers lack confidence in their choices and express fear towards adverse health events and unanticipated costs. **Joseph Ejiyi et al.(2018)**examined an insurance dataset from the Zindi Africa competition, which was reported to be from Olusola Insurance Company in Lagos, Nigeria, to illustrate the effectiveness of each of the ML algorithms we used here. The findings indicated that, based on a dataset received from Zindi, insurance authorities, shareholders, administration, finance experts, banks, accountants, insurers, and customers all showed concern regarding insurance company insolvency. This worry stemmed from a perceived requirement to shield the general public from the repercussions of insurer insolvencies while also lowering management and auditing duties. In this work, we offer a strategy for preventing insurance company insolvency. **Fauzan**

and Murfi (2018) used XGBoost to solve the issue of claim prediction and evaluate its accuracy. They also compared XGBoost's performance against that of other ensemble learning methods, such as AdaBoost, Stochastic GB, Random Forest, and Neural Network, as well as online learning methods. In terms of normalised Gini, the simulations suggest that XGBoost outperforms other techniques. People are increasingly investing in such insurance, allowing con artists to defraud them. Insurance fraud is a crime that can be committed by either the customer or the insurance contract's vendor. Unrealistic claims and post-dated policies, among other things, are examples of client-side insurance fraud. Insurance fraud occurs on the vendor side in the implementation of regulations from non-existent firms and failure to submit premiums, among other things.

Objectives

- To identify the factors determining the premium
- To understand the corelation between the factors and the premium rates
- To build a model to predict the health insurance premium

Research Methodology

This study relies on secondary data, with independent variables like age, medical history, and lifestyle factors influencing the dependent variable—premium price. Data analysis begins with descriptive statistics, followed by exploratory data analysis (EDA) using MS Excel and Python to identify patterns and relationships. For accurate premium forecasting, predictive models such as Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors (KNN) are applied to capture complex relationships and improve estimation accuracy.

Exploratory Data Analysis

Fig.1

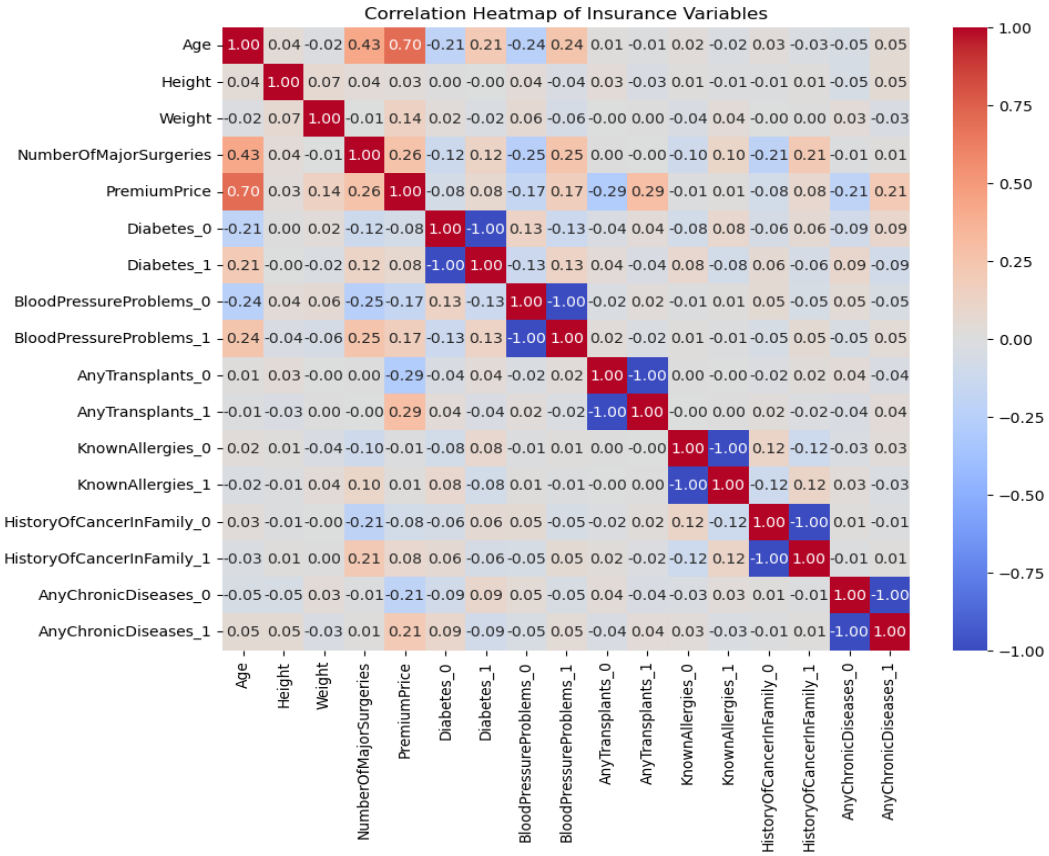
Table showing different measures of the data

	Age	Diabetes	BloodPressureProblems	AnyTransplants	AnyChronicDiseases	Height	Weight	KnownAllergies	HistoryOfCancerInFamily	NumberOfMajorSurge
0	45	0	0	0	0	155	57	0	0	
1	60	1	0	0	0	180	73	0	0	
2	36	1	1	0	0	158	59	0	0	
3	52	1	1	0	1	183	93	0	0	
4	38	0	0	0	1	166	88	0	0	

The above table shows a clear picture of the data collected and used for analysis.

Fig.2

Heat Map showing the correlations between the variables



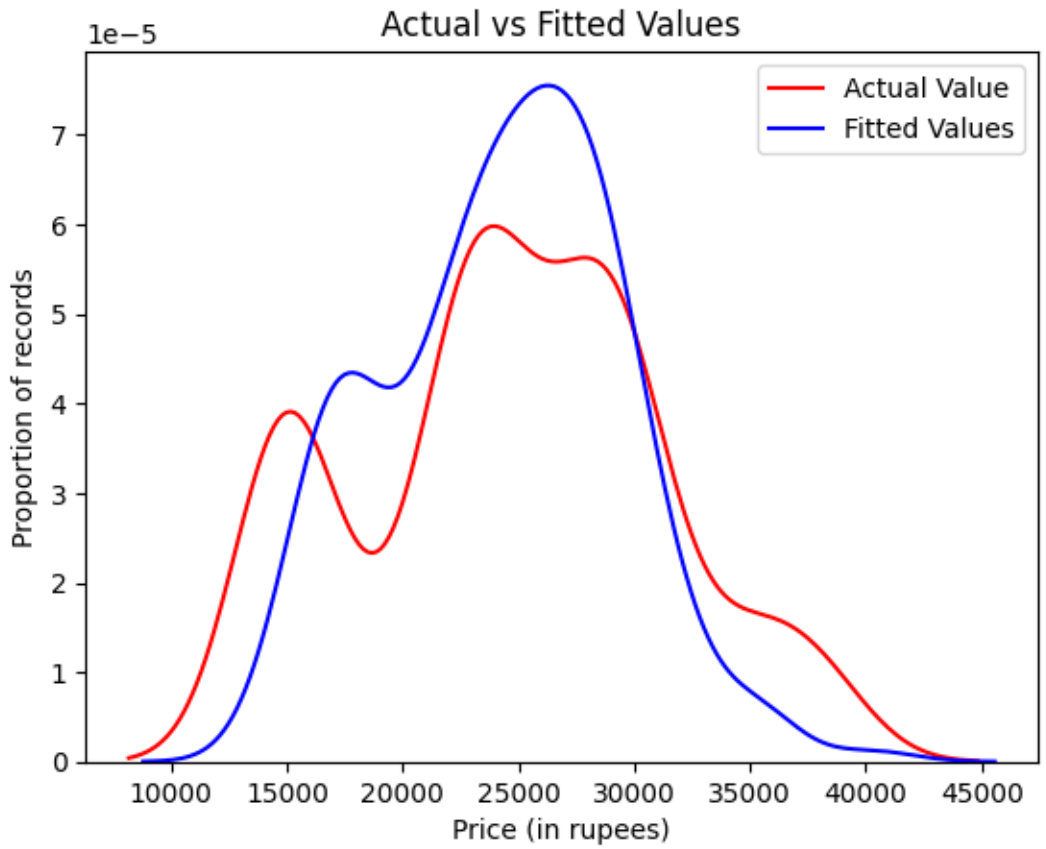
The heat map in Fig 2 shows the values of correlation, hence proving that premium price and Age have a high correlation of 0.70.

Predictive Analysis:

Models like Multiple Linear Regression , Logistic Regression ,Random Forest Regressor, Gradient Booster Regressor and Support Vector Machine were used to predict the Premium price.

Multiple Linear Regression:

Plot showing the actual and predicted values of the Premium Price

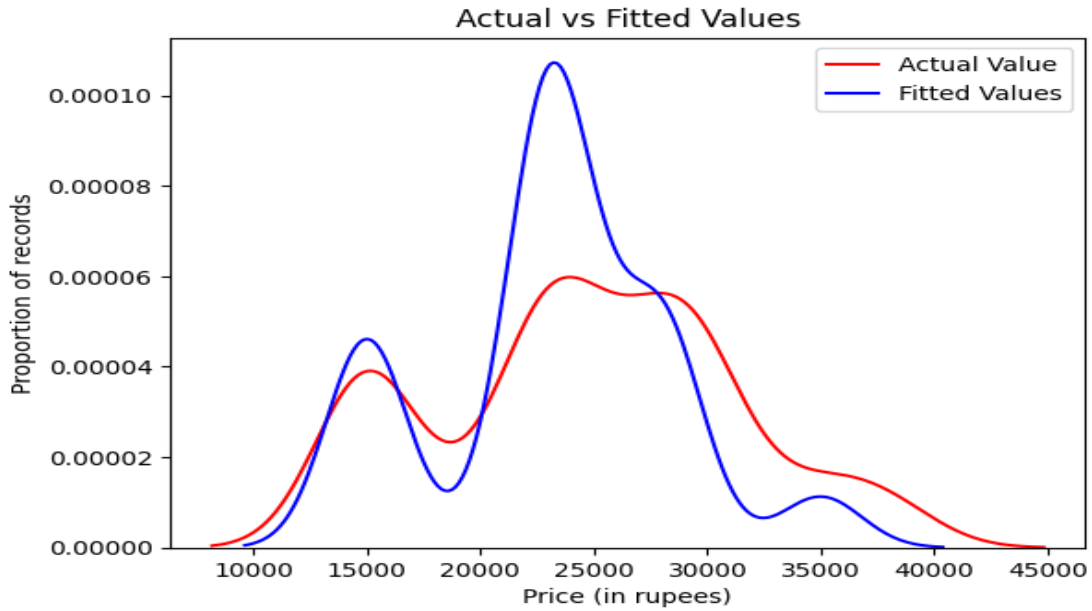


Linear Regression RMSE: 3501.6274856032987

Linear Regression R2: 0.7124626554287766

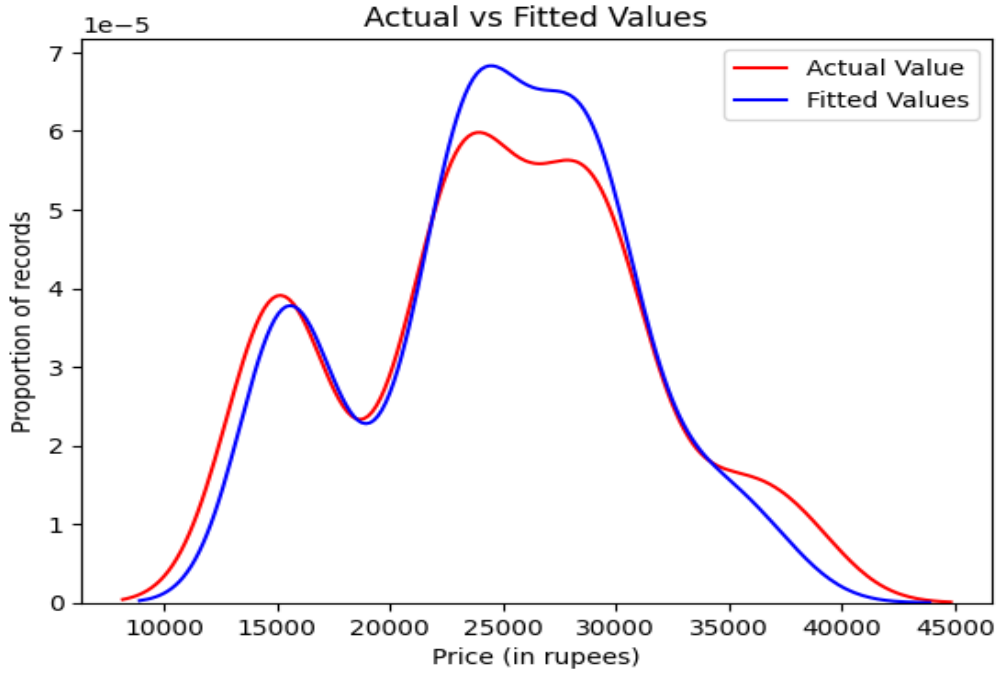
Logistic Regression:

Plot showing the actual and predicted values of the Premium Price



Logistic Regression Accuracy: 0.5808080808080808

Random Forest Regressor:

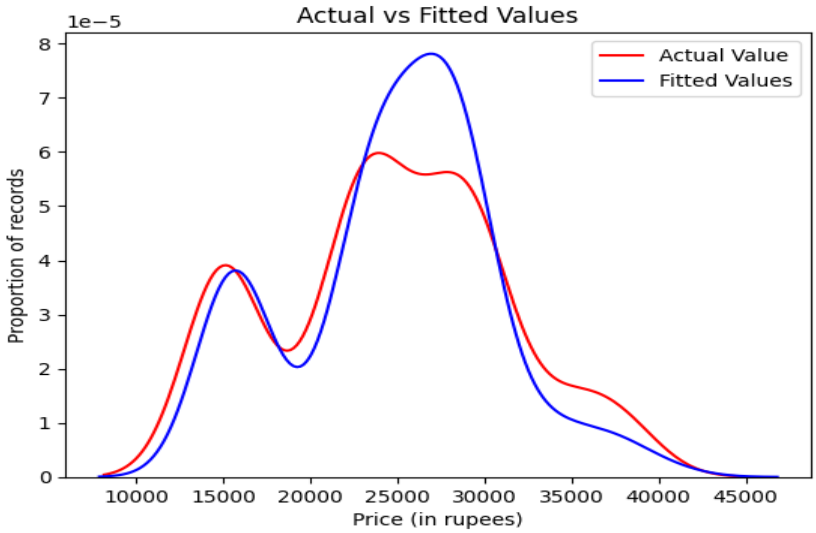


Random Forest RMSE: 2247.41377740475

Random Forest R2: 0.8815539859597743

Gradient Booster Regressor:

Plot showing the actual and predicted values of the Premium Price



GBM - Mean Squared Error: 6135564.4469934795

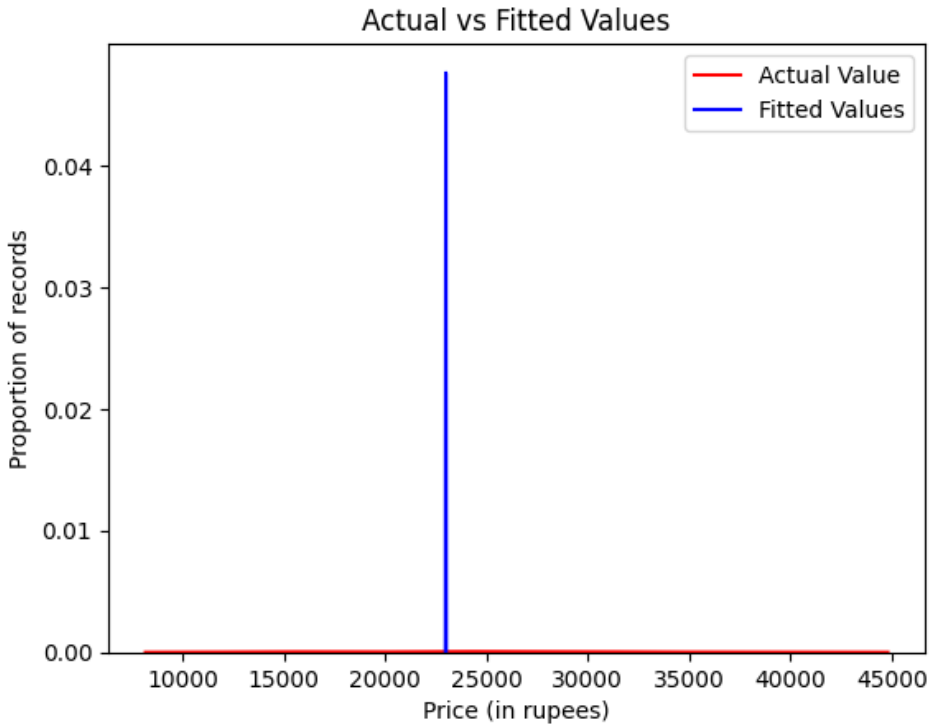
GBM - Root Mean Squared Error: 2477.0071552164477

GBM - R^2 Score: 0.8561171953405422

Support Vector Machine:

Another basic method that any machine learning expert should have in his or her arsenal is a support vector machine. Many people favour support vector machines because they generate substantial accuracy while using minimal compute power. SVM, or Support Vector Machine, may be used for both regression and classification applications. However, it is commonly employed in classification aims

Plot showing the actual and predicted values of the Premium Price



SVM - Mean Squared Error: 45132047.42818556

SVM - Root Mean Squared Error: 6718.038957030955

SVM - R^2 Score: -0.05837459945075296

Key Findings

Age is a crucial determinant in premium estimation, with a strong correlation of 0.70. Older individuals tend to have higher premiums due to increased health risks and medical expenses. Chronic diseases also significantly impact premium costs, as conditions like diabetes, hypertension, and heart disease lead to higher medical expenditures, increasing financial risk for insurers. The amount of major surgeries one has had also impacts premiums, with a history of surgeries showing potential recurring health problems and medical expenses down the line. Lifestyle elements like weight and known allergies also factor into premium differences, as these influence overall health risk. A personal history of cancer also factors in, as genetic makeup can result in increased insurance rates.

Conclusion

Health insurance premium estimation necessitates the correct evaluation of risk to determine equitable pricing and financial viability. Variables like age, chronic conditions, previous operations, and hereditary factors all add to variations in premiums, which necessitate predictive modeling. Machine learning models, especially Random Forest Regression, enhance precision in premium estimation through the effective examination of these variables. Through the application of advanced predictive methods, insurers can streamline pricing models and provide customized insurance plans. Future studies can incorporate additional variables like lifestyle habits and claim histories to further enhance predictive accuracy and transparency.

Bibliography and References:

1. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9265373/

2. ul Hassan C.A., Iqbal J., Hussain S., AlSalman H., Mosleh M.A.A., Sajid Ullah S. A Computational Intelligence Approach for Predicting Medical Insurance Cost. *Math. Probl. Eng.* 2021;**2021**:1162553. doi: 10.1155/2021/1162553.

A Computational Intelligence Approach for Predicting Medical Insurance Cost (hindawi.com)

3. Cevolini A., Esposito E. From Pool to Profile: Social Consequences of Algorithmic Prediction in Insurance. *Big Data Soc.* 2020;**7** doi: 10.1177/2053951720939228.

From pool to profile: Social consequences of algorithmic prediction in insurance - Alberto Cevolini, Elena Esposito, 2020 (sagepub.com)

4. van den Broek-Altenburg E.M., Atherly A.J. Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season. *Appl. Sci.* 2019;**9**:2035. doi: 10.3390/app9102035.

Applied Sciences | Free Full-Text | Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season (mdpi.com)

5. Hanafy M., Mahmoud O.M.A. Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models. *Int. J. Innov. Technol. Explor. Eng.* 2021;**10**:137–143. doi: 10.35940/ijitee.C8364.0110321.

C83640110321 - International Journal of Innovative Technology and Exploring Engineering (IJITEE)

6. Bhardwaj N., Anand R. Health Insurance Amount Prediction. *Int. J. Eng. Res.* 2020;**9**:1008–1011. doi: 10.17577/IJERTV9IS050700.

https://doi.org/10.17577%2FIJERTV9IS050700

7. Boodhun N., Jayabalan M. Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. *Complex Intell. Syst.* 2018;**4**:145–154. doi: 10.1007/s40747-018-0072-1.CrossRefRisk prediction in life insurance industry using supervised learning algorithms | Complex & Intelligent Systems (springer.com)

8. Goundar S., Prakash S., Sadal P., Bhardwaj A. Health Insurance Claim Prediction Using Artificial Neural Networks. *Int. J. Syst. Dyn. Appl.* 2020;**9**:40–57. doi: 10.4018/IJSDA.2020070103.

Health Insurance Claim Prediction Using Artificial Neural Networks | IGI Global (igi-global.com)

9. Ejiyi C.J., Qin Z., Salako A.A., Happy M.N., Nneji G.U., Ukwuoma C.C., Chikwendu I.A., Gen J. Comparative Analysis of Building Insurance Prediction Using Some Machine Learning Algorithms. *Int. J. Interact. Multimed. Artif. Intell.* 2022;**7**:75–85. doi: 10.9781/ijimai.2022.02.005.

ijimai7_3_7.pdf

10. Rustam Z., Yaurita F. Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means. *J. Phys. Conf. Ser.* 2018;**1028**:012118. doi: 10.1088/1742-6596/1028/1/012118.

Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means - IOPscience