IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research Paper , Journal Volume 10, Iss 10, 2021

IMPLEMENTATION OF PREDICTING MEDICAL INSURANCE COST WITH MACHINE LEARNING TECHNIQUES

[1]Shaik Babjan,[2]P Nagaraju,[3]Pula Sekhar,[4]Badaghar Javeed Basha

[1]Assistant Professor,[2,3,4]Associate Professor
Department of CSE

Tadipatri Engineering College,Tadipatri,AP

ABSTRACT:
With health insurance, the financial risk of unexpected medical costs is distributed among many people, reducing the amount of money at risk. Global public health expenditure has almost quadrupled over the last 20 years, and in 2023 it is expected to reach $8.5 trillion, or 9.8% of the world GDP after accounting for inflation. Multinational multi-private sectors supply 70% of outpatient care and 60% of complete medical treatments, sometimes at excessive costs. Health insurance has become a necessary good due to rising healthcare costs, higher life expectancies, and the rise of non-communicable illnesses. The expansion of insurance data availability has made it possible for insurance companies to use predictive modelling to improve their customer service and corporate operations. In order to forecast future output values based on consumer behaviour patterns, insurance policies, data-driven decision-making, and the creation of new schemes, historical insurance data is examined using computer algorithms and machine learning (ML). The insurance business has shown great promise for machine learning (ML), which led to the creation of the ML Health Insurance Prediction System. This cost-price prediction system makes it easier to determine premium values quickly and efficiently, which lowers medical costs. Three regression models are compared and contrasted in this system: Random Forest Regressor, Support Vector Regression, and Linear Regression. The models were trained on a dataset, which allowed for the generation of predictions and the validation of the model's efficacy by comparison with real data.
Keywords: Random forest regression, Support Vector Machine, Health insurance prediction, machine learning, and linear regression
.

1. INTRODUCTION

General insurance plays a vital role in protecting individuals and their valuable assets, such as homes, vehicles, and real estate, from unforeseen events and accidents. It offers coverage against a range of risks, including fire accidents, earthquakes, floods, thefts, storms, travel accidents, and legal liabilities. Amongst these, health insurance holds particular importance as it ensures a secure and stable life by safeguarding against unexpected medical expenses that can disrupt financial stability and long-term goals[5]. Given the complexities of modern health challenges, planning for healthcare has become a necessity, leading to the availability of insurance plans for individuals and families.

In India, a significant proportion (around 75%) of the population currently bears their medical expenses out of pocket. However, health insurance coverage has been increasing steadily, with approximately 514 million people covered during the fiscal year 2021. According to the NITI Aayog Health Index 2021, Kerala has been ranked as the healthiest state in India, with a composite score of 82.90. The insurance industry in India comprises 57 firms, including 33non-life insurers and 24 life insurers, with seven public sector companies playing a prominent role. Strong competitors have also emerged in the form of private insurers such as ICICI, HDFC, SBI, and Star Health[7].

Previous studies have shown that individuals enrolled in Medicare tend to have more favorable assessments of their insurance compared to those with commercial plans. Various studies have compared Medicaid and commercial insurance,but the findings have been conflicting and limited to specific populations or service utilization. Recent data explicitly comparing the experiences of individuals with public and private health insurance is lacking[3].

The objective of this Paper is to provide accurate estimatesof health insurance costs for different providers andindividuals. While predictions may not always follow a consistent pattern, they can assist in making informed decisions regarding the selection of appropriate health insurance coverage[8]. Early cost calculations can help individuals evaluate their options more carefully and ensure they choose the most suitable coverage. Furthermore, the research may offer insights into maximizing the benefits of health insurance.

2. LITERATURE SURVEY

India's market for general insurance is growing significantly in the post-liberalization environment. The opening of the Indian insurance market to foreign companies, Third Party Administrators, low insurance premiums, quick and immediate settlement of insurance claims, innovative general insurance policies, discounts on insurance products, growing public awareness, more distribution channels, and other factors have all contributed to this market's spectacular growth.The Below includes various research papers and articles related to different aspects of health insurance. [1]. "Operational Efficiency of Selected General Insurance Companies in India" - This paper explores the operational efficiency of general insurance companies in India, particularly in the context of competition between public and

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research Paper **Journal Volume 10, Iss 10, 2021**

private insurers.[2]. "An Empirical Evaluation On Proclivity Of Customers Towards Health Insurance During Pandemic" - The research focuses on studying the awareness and inclination of the public towards health insurance during a pandemic, using SPSS software for analysis.[3]. "Health Insurance in India - An Overview" - This article provides an overview of the health insurance industry in India, including the growth and development of standalone health insurers and government-sponsored health insurance providers.[4]. "A Conceptual Review Paper on Health Insurance in India" - The paper reviews existing literature on health insurance in India to understand the growth and potential benefits of health insurance for the population.
[5]. "Need-based and Optimized Health Insurance Package Using Clustering Algorithm" - This research proposes the use of clustering algorithms to design health insurance packages based on the specific needs of employees, aiming to provide optimized coverage.[6]. "Health Insurance Amount Prediction" - The authors analyze personal health data to predict insurance amounts for individuals using regression models. Multiple Linear Regression and Gradient Boosting Decision Tree Regression are compared for their performance.[7]. "Predicting the Risk of Disease Using Machine Learning Algorithm" - The study aims to predict the risk of chronic kidney disease (CKD) using machine learning algorithms, specifically by building a regression model to predict creatinine values and combining them with other health-related features.[8]. "Piecewise-linear Approach for Medical Insurance Costs Prediction Using SGTM Neural-like Structure" - This article proposes a method for predicting medical insurance costs using a piecewise-linear approach and the SGTM neural-like structure, comparing it with other methods like multilayer perceptron.[9]. "Predicting Health Care Costs Using Evidence Regression" - The research investigates the use of an interpretable regression method based on the Dempster-Shafer theory, called Evidence Regression, for predicting health care costs. It outperforms Artificial Neural Network and Gradient Boosting methods in terms of accuracy.[10]. "Health Insurance Sector in India: An Analysis of Its Performance" - This study analyzes the performance of the health insurance sector in India, specifically examining the relationship between premium earnings and underwriting loss using regression analysis. [11]. "Knowledge and Understanding of Health Insurance" - The research focuses on health insurance literacy and disparities in knowledge among different socioeconomic groups in Israel, emphasizing the need for tailored communication strategies and simplified plan information. [12]. "The Effects of Health Insurance on Health-Seeking Behaviour: Evidence from the Kingdom of Saudi Arabia" - The study explores the impact of health insurance on health-seeking behavior in Saudi Arabia and suggests the introduction of national health insurance coverage as an effective measure to improve access to healthcare.

3. PROPOSED MEDICAL HEALTH INSURANCE COST PREDICTION SYSTEM

The dataset used here contains information related to health insurance costs and various factors that influence them. The dataset has 7 columns and 1338 rows.
Based on prediction ,we can identify some of the important columns/features in the dataset:
1. Age: Represents the age of the insured individual.
2. Smoking Status: Indicates whether the insured individual is a smoker or a non-smoker.
3. BMI: Represents the Body Mass Index, a measure of body fat based on height and weight.
4.No. of Childrens: Provides information about the insured children's count.
5.Sex: Indicates the Gender.
6. Region : Represents the geographical region of the insured individual.
7. Charges: Represents the medical insurance charges or costs.
To predict the cost of health insurance, the dataset needs to be cleaned and prepared before applying regression algorithms. The information suggests that age and smoking status have the most significant impact on insurance costs, with smoking having the greatest effect. Other factors such as No. of Children's , BMI, marital status, and geography also play a role in determining insurance costs.

S. no	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86
5	31	female	25.7	0	no	southeast	3756.62
6	46	female	33.4	1	no	southeast	8240.59
7	37	female	27.7	3	no	northwest	7281.51
8	37	male	29.8	2	no	northeast	6406.41
9	60	female	25.8	0	no	northwest	28923.14

Fig 1. Data Set

3.1 TECHNOLOGY USED:

A. Machine Learning:

Machine learning is a branch of artificial intelligence that concentrates on algorithms and models enabling computers to learn from data, make predictions, or make decisions without requiring explicit programming. It involves training models on historical data and using them to make predictions or classify new, unseen data based on patterns and relationships learned during training.

B. SVM (Support Vector Machines):

SVM is a supervised machine learning algorithm used for both classification and regression tasks. It works by finding an optimal hyperplane that separates different classes in a high-dimensional feature space.SVM aims to maximize the margin (distance) between the decision boundary and the data points of different classes, allowing for better generalization and improved performance on unseen data.It can handle

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 10, Iss 10, 2021

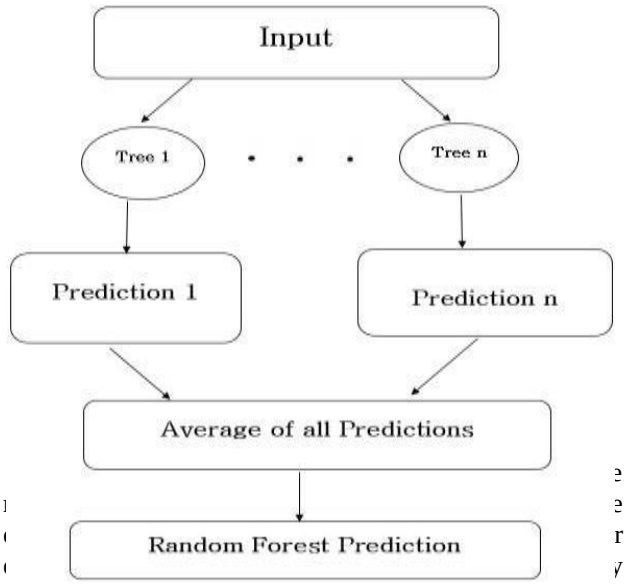
linear and non-linear classification problems using different kernel functions, such as linear, polynomial, or radial basis function (RBF).

C. Random Forest:

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions.It is a

A supervised learning algorithm is commonly employed for both classification and regression tasks. Random Forest builds an ensemble of decision trees bytraining each tree on a randomly selected subset of features and data samples.

During prediction, each tree in the forest independently makes a prediction, and the final prediction is determined based on a majority vote (for classification) or averaging (for regression) of the individual tree Predictions. Random Forest is known for its ability to handle high-dimensional data, provide feature importance estimates, and handle non-linear relationships between features and the target variable.



the optimal line of best fit that minimizes the disparity between the predicted values and the actual values. It assumes a linear relationship between the input features and the target variable. Linear regression can be extended to handle multiple variables (multiple linearregression) or non-linear relationships by using polynomial or other non-linear transformations of the input features.

4.RESULT

The proposed system's dataset was tested with three machine learning algorithms: Random Forest, Linear Regression, andSupport Vector Regressor. The accuracy of each algorithm was measured, and the results are as follows:

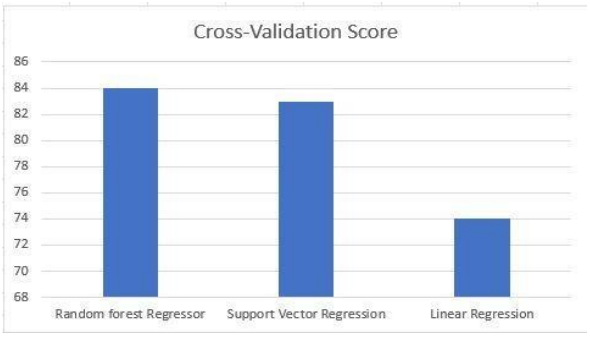


Fig 3. Performance Graph of Proposed System with Three ML Algorithms

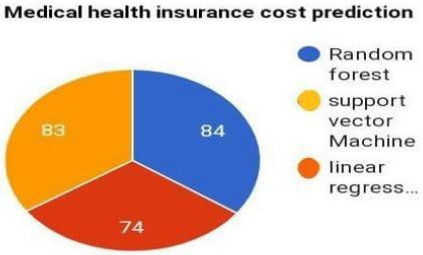


Fig 2. Random Forest

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$$

1. Random Forest: 84% accuracy
2. Linear Regression: 74% accuracy
3. Support Vector Regressor: 83% accuracy

These accuracy percentages indicate how well the algorithmsperformed in predicting the target variable based on the givendataset. It seems that Random Forest achieved the highest accuracy of 84%, followed by Support Vector Regressor with83% accuracy, and Linear Regression with 74% accuracy Shown in fig 3.

5.CONCLUSION AND FUTURE SCOPE

It seems that you are summarising the conclusions and possible uses of the regression models created using data from health insurance policies. Of the three models examined, the random forest regression model fared the best, according to your assertion. Across all algorithms, age and smoking status were shown to be the most significant determinants of insurance rates. Irrelevant qualities were eliminated from the features and various attribute combinations were investigated in order to increase accuracy. It's possible that this procedure improved the models' prediction power and refinement.

FUTURE SCOPE

When compared to other algorithms, the Random Forest

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 10, Iss 10, 2021

algorithm's unpredictability may result in better prediction accuracy. It is advised to test the system in the future on a dataset with at least a million entries in order to evaluate its scalability. Large scale data processing like this calls for distributed frameworks like Hadoop and Spark. These frameworks are designed to process and distribute data over many cluster nodes, facilitating parallel computing and improving scalability.Because of its scalability, the system can handle large volumes of data while still operating efficiently.

REFERENCES

1.Bc Lakshmana , P.Jayarami Reddy, P.Sravan Kumar ”Operational Efficiency of Selected General Insurance Companies in India” (2019) .

2. SatakshiChatterjee,Dr.Arunangshu,Dr.S.N. .Bandyopadhyay “An Empirical Evaluation On Proclivity Of Customers Towards Health Insurance “(2018)

3. K Swathi and R Anuradha ,” Health insurance in India”(2017) https://www.iosrjournals.org/iosr-jbm/papers/Conf.17037-2017/Volume-7/10.%2049-52.pdf

4. Dr. Vazir Singh Nehra, Suman Devi, “A Conceptual Review Paper On Health Insurance in India”(2017)

5. Matloob I, Khan SA, Hussain F, Butt WH, Rukaiya R, Khalique F (2021) Need-based and optimized health insurance package using clustering algorithm. Appl Sci 11(18):8478. https://doi.org/10.3390/app11188478

6. Bhardwaj N, Delhi RA, Akhilesh ID, Gupta D (2021) Health insurance amount prediction [Online]. https://economictimes.indiatimes.com/wealth/insure/what-you-need-to

7. Wang W, Chakraborty G, Chakraborty B (2021) Predicting the risk of chronic kidney disease (CKD) using machine learning algorithm .appl sci 11(1):1–17. https://doi.org/10.3390/app11010202

8. Tkachenko R, Izonin I, Kryvinska N, Chopyak V, Lotoshynska N, Danylyuk D (2018) Piecewise-linear approach for medical insurance costs prediction using SGTM neural-like structure. CEUR Workshop Proc 2255:170–179

9. Panay B, Baloian N, Pino J, Peñafiel S, Sanson H, Bersano N (2019) Predicting health care costs using evidence regression. Proceedings (1):74. https://doi.org/10.3390/proceedings2019031074

10. Binny, Dr. Meenu Gupta “Health insurance in India-Opportunities and challenges”(2017)

11. Dutta, M.M, "Health insurance sector in India: an analysis of its performance", Vilakshan - XIMB Journal of Management 2020, Vol. 17 No. 1/2, pp. 97-109.

12. Barnes, A.J., Hanoch, Y. Knowledge and understanding of health insurance: challenges and remedies. Barnes and Hanoch Israel Journal of Health Policy Research 2017.