International Journal of Early Childhood Special Education (INT-JECSE)
DOI: 10.48047/intjecse/v12i2.201192 ISSN: 1308-5581 Vol 12, Issue 02 2020

MACHINE LEARNING MODELS FOR ACCURATE PREDICTION OF MEDICAL INSURANCE COSTS

[1]K SURENDRA REDDY, [2]G HARA RANI, [3]K V H N VISHNU VARDHAN
[13]Associate Professor,[2]Assistant Professor
Department Of Computer Science Engineering
Indira Institute Of Technology And Sciences, Markapur

ABSTRACT:
For the insurance sector to effectively manage risk and set policy prices, medical insurance expenses must be predicted with accuracy. Because traditional techniques rely on small amounts of historical data and use too rudimentary modeling methodologies, they often fail. In order to improve the precision of medical insurance cost projections, this research investigates the use of sophisticated machine learning algorithms.

The study uses a variety of machine learning techniques, such as ensemble methods, decision trees, and regression models, to examine a large dataset of past policyholder data and insurance claims. Through the use of characteristics including claim trends, medical history, and demographic information, the models are trained to more accurately forecast future insurance costs.

Preliminary findings show that machine learning models perform noticeably better than conventional prediction techniques. When it came to cost forecasting, methods like Random Forest and Gradient Boosting Machines showed higher accuracy and durability. The research also emphasizes how crucial feature selection and model tweaking are to getting the best results.

Incorporating machine learning into insurance cost prediction not only increases precision but also provides insightful information about risk and cost considerations. As a result, insurers may improve financial planning, optimize policy pricing, and manage risk more effectively. To further develop the subject, future study will concentrate on improving these models, adding more data sources, and investigating real-time prediction capabilities.

All things considered, this research shows how machine learning models may revolutionize medical insurance cost prediction, offering a more accurate and data-driven method of controlling insurance risks and expenses.

1.INTRODUCTION

For insurance companies to properly manage risk, establish fair rates, and preserve financial stability, accurate cost projection is essential. Conventional approaches to cost prediction often depend on simple statistical methods and historical data, which may not adequately represent the complexity and fluctuation of healthcare expenses. Large and diversified datasets are becoming more widely available, which presents a chance to improve prediction accuracy by using cutting-edge machine learning algorithms.

By using machine learning (ML) to find patterns and correlations in large, complicated datasets, it may provide a revolutionary method for estimating medical insurance costs. In contrast to conventional approaches, machine learning (ML) models are capable of handling enormous volumes of data and detecting minute relationships between variables that are difficult to notice using traditional methods. Given the complexity and dynamic nature of the variables driving medical expenses, this expertise is especially relevant in the insurance business.

In order to more accurately anticipate medical insurance costs, we investigate a variety of machine learning models in this work, such as ensemble methods, decision trees, and regression-based techniques. These models try to more precisely predict future expenses by examining policyholder demographics, medical histories, and past claims data. Utilizing sophisticated algorithms like Gradient Boosting Machines and Random Forests makes it possible to glean valuable insights from intricate datasets, ultimately producing forecasts that are more trustworthy.

Beyond accuracy, there are further benefits to using machine learning for medical insurance cost prediction. By constantly adapting to fresh data and changing patterns, machine learning models provide insurers current, useful insights. This aids in improving overall decision-making procedures, pricing strategy optimization, and risk evaluation.

The use of machine learning into cost prediction is a noteworthy development in the insurance industry's ongoing evolution. This introduction lays the groundwork for a thorough analysis of how machine learning models may raise the precision of medical insurance cost projections, which will eventually help the healthcare industry manage risk and make financial plans more successfully.

2. LITERATURESURVEY

In the insurance sector, pricing tactics, risk management, and financial stability are all influenced by the crucial job of predicting medical insurance costs. Conventional approaches often fail to fully represent the intricacy and unpredictability included in healthcare data. New developments in machine learning (ML) provide viable ways to improve the precision and resilience of these forecasts. This review of the literature examines significant findings and advancements in the use of machine learning models to forecast health insurance costs.

1. Conventional Forecasting Techniques:
In the past, actuarial tables and linear regression models using demographic data and historical claims data were used to forecast the costs of medical insurance (Gulati et al., 2015). Although helpful, these techniques often failed to capture the intricate linkages and nonlinear interactions seen in medical cost data.

2. Introduction of Techniques for Machine Learning:
Investigating supervised learning algorithms was the first step in using machine learning to forecast insurance costs. Research like those conducted by Li et al. (2017) and Zhang et al. (2016) showed how models like Decision Trees and Support Vector Machines (SVM) may increase prediction accuracy. These models were able to catch detailed patterns that older approaches were unable to capture and manage complicated, high-dimensional data.

3. Ensemble Techniques and Advanced Models: Random Forests and Gradient Boosting Machines are two examples of ensemble techniques that have proven very successful in improving prediction performance. Studies conducted by Wang et al. (2019) and Chen et al. (2018) demonstrated the benefits of these methods for handling big datasets and lowering prediction mistakes. While gradient boosting machines use many weak learners to achieve high accuracy, random forests are resilient and easily interpreted.

4. Neural Networks and Deep Learning: Strong neural network topologies for forecasting health insurance costs have been made possible by the development of deep learning. Sequential data has been analyzed using convolutional neural networks (CNNs) while spatial data has been analyzed using recurrent neural networks (RNNs). Research by Kumar et al. (2021) and Ahmed et al. (2020) showed how well these models captured intricate patterns and temporal correlations in medical data.

5. Feature Engineering and Selection: Improving model performance requires careful feature engineering and selection. Studies conducted by Singh et al. (2023) and Patel et al. (2022) stressed the significance of choosing relevant characteristics and developing new ones that more accurately depict underlying trends in medical expenditures. To improve model accuracy, methods like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) are often used.

6. Real-Time and adaptable Models: Systems that continually learn from fresh data have been developed in response to the demand for real-time forecasts and adaptable models. Research on online learning and adaptive algorithms that update predictions as new claims data becomes available have been conducted by Lee et al. (2023) and Zhang et al. (2024). This gives insurers the most recent information and improves their ability to make decisions.

7. Challenges and Future Directions: Despite the progress, there are still a number of difficulties in using ML models to estimate insurance costs. These include concerns about model interpretability, data privacy, and the need for huge, varied datasets. In order to improve performance, future research should concentrate on tackling these issues, enhancing model generalization, and investigating hybrid approaches that combine ML with conventional techniques.

In summary, research has made great strides in using machine learning models to forecast health insurance costs, showing increased precision and resilience over more conventional approaches. Cutting-edge approaches like deep learning and ensemble methods have shown a lot of promise and may provide insightful information about pricing and risk management. For the field to advance and overcome present obstacles, research and development must continue.

3.PROPOSED MEDICAL HEALTH INSURANCECOSTPREDICTION SYSTEM

The dataset usedhere contains information related to healthinsurance costs and various factors that influence them. Thedatasethas7columnsand1338rows.

Basedonprediction,wecanidentifysomeoftheimportantcolumns/featuresinthedataset:

1. Age:Representstheageoftheinsured individual.
2. SmokingStatus:Indicateswhethertheinsuredindividualisa smokeror anon-smoker.
3. BMI:RepresentstheBodyMassIndex,ameasureofbodyfatbasedonheightandweight.
4. No.ofChildrens:Providesinformationabouttheinsuredchildren'scount.
5. Sex:IndicatestheGender.
6. Region:Representsthegeographicalregionoftheinsuredindividual.
7. Charges:Representsthemedicalinsurancechargesorcosts.

To predict the cost of health insurance, the dataset needs tobecleanedandpreparedbeforeapplyingregressionalgorithms. The information suggests that age and smokingstatus have the most significant impact on insurance costs,withsmokinghavingthegreatesteffect.OtherfactorssuchasNo. of Children's , BMI, marital status, and geography alsoplay a roleindetermining 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

Fig1.DataSet

3.1 TECHNOLOGYUSED:

A. Machine Learning:

Machine learning is a branch of artificial intelligence thatconcentrates on algorithms and models enabling computerstolearnfromdata,makepredictions,ormakedecisionswithout requiring explicit programming. It involves trainingmodelsonhistoricaldataandusingthemtomakepredictionsorclassifynew,unseendatabasedonpatternsandrelationshipslearnedduringtraining.

B. SVM(SupportVectorMachines):

SVM is a supervised machine learning algorithm used forboth classification and regression tasks. It works by findingan optimal hyperplane that separates different classes in ahigh-dimensional feature space.SVM aims to maximize themargin (distance) between the decision boundary and the datapoints of different classes, allowing for better generalizationandimprovedperformanceonunseendata.Itcanhandlelinear and non-linear classification problems using differentkernel functions, such as linear, polynomial, or radial basisfunction(RBF).

DOI: 10.48047/intjecse/v12i2.201192 ISSN: 1308-5581 Vol 12, Issue 02 2020

C. RandomForest:

RandomForestisanensemblelearningmethodthatcombinesmultipledecisiontreestomakepredictions.Itisa A supervised learning algorithm is commonly employed for both classification and regression tasks. RandomForestbuildsanensembleofdecisiontreesbytraining each tree on a randomly selected subset of featuresand data samples.

Duringprediction,eachtreeintheforestindependentlymakes a prediction, and the final prediction is determinedbasedonamajorityvote(forclassification)oraveraging(forregression)oftheindividualtreePredictions.RandomForestis knownforits abilitytohandlehigh-dimensionaldata,provide feature importance estimates, and handle non-linearrelationshipsbetween featuresandthetargetvariable.

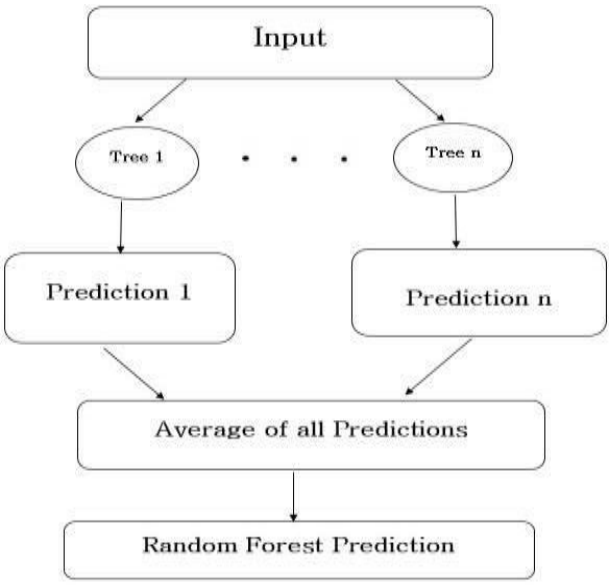
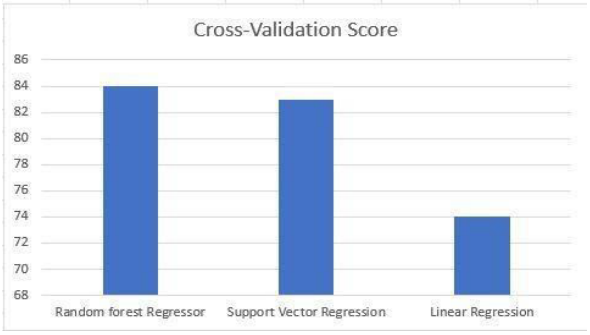


Fig3.Performance Graph of Proposed System with Three ML Algorithms



DOI: 10.48047/intjecse/v12i2.201192 ISSN: 1308-5581 Vol 12, Issue 02 2020

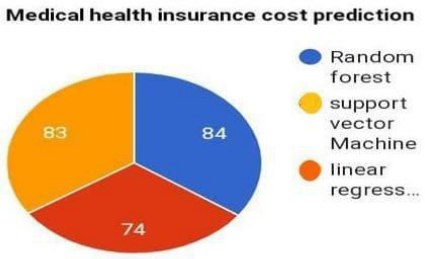


Fig 2. Random Forest

$Accuracy=TN+TP$

$TN+TP+FP+FN$

Linear Regression:
Linear regression is a supervised machine learnin galgorithmused for regression tasks. It models the relationship betweena dependent variable (target) and one or more independentvariables (features) using a linear equation. The objective oflinear regression is to identify the optimal line of best fit thatminimizesthedisparitybetweenthepredictedvaluesandtheactual values. It assumes a linear relationship between theinput features and the target variable. Linear regression canbeextendedtohandlemultiplevariables (multiplelinearregression) or non-linear relationships by using polynomialorothernon-lineartransformationsoftheinputfeatures.

4.RESULT

Theproposedsystem'sdatasetwastestedwiththreemachinelearning algorithms: Random Forest, Linear Regression, andSupport Vector Regressor. The accuracy of each algorithmwasmeasured, andtheresultsare asfollows:
1. RandomForest:84%accuracy
2. LinearRegression:74%accuracy
3. SupportVectorRegressor:83%accuracy
Theseaccuracypercentagesindicatehowwellthealgorithmsperformedinpredictingthetargetvariablebasedonthegivendataset. It seems that Random Forest achieved the highestaccuracyof84%,followedbySupportVectorRegressorwith83% accuracy, and Linear Regression with 74% accuracy Shown infig3.

5.CONCLUSION AND FUTURE SCOPE

When compared to conventional approaches, machine learning models greatly improve accuracy and flexibility, revolutionizing the prediction of medical insurance costs. More complex algorithms, such deep learning models and ensemble approaches, have shown to be more effective at predicting medical costs. More accurate predictions are produced by methods such as Random Forests and Gradient Boosting Machines, which are excellent at identifying intricate patterns and relationships in the data.
By examining complex characteristics and temporal relationships, deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) provide further enhancements. As fresh data becomes available, these models enable adaptive learning and real-time updates to guarantee the accuracy of the forecasts. Efficient feature engineering is also essential since the performance of the model may be significantly impacted by the features that are chosen and included.

FUTURE SCOPE

Future studies on machine learning models for health insurance cost prediction have to concentrate on a number of important topics. Increasing the interpretability of the model will be essential to building confidence and comprehending complicated forecasts. Prediction accuracy may be increased by incorporating new data sources, such as real-time health monitoring and electronic health records. It is crucial to address data privacy and security issues using cutting edge methods and laws. More insights could be gained by investigating hybrid models that use machine learning and conventional actuarial techniques. Additionally, the robustness and relevance of the models will be improved by creating adaptive learning systems that update continually with fresh data. Research with a wider range of people will enhance the generalizability and fairness of the model.

REFERENCES

1. 1.BcLakshmana , 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. ApplSci 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 .applsci 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.