By contrasting decision trees with logistic regression, a novel categorization-based cost prediction method for health insurance may be developed under supervision.

G.Satya Mounika Kalyani [1], Deepa.N [2]*

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, India, PinCode: 602105

[2]*Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, India, PinCode: 602105

ABSTRACT

AIM: The aim of research is to classify the cost prediction in health insurance using novel ranking with machine learning algorithms. **Materials and Methods**: The categorizing is performed by adopting a sample size of n=10 in Decision Tree and sample size n = 10 in Logistic Regression algorithms was iterated 20 times for efficient and accurate analysis on labeled images with G power in 80% and threshold 0.05%, CI 95% mean and standard deviation. **Results**: The analysis of the results shows that the Decision Tree has a high accuracy of (94.30%) in comparison with the Logistic Regression algorithm (92.70%). There is a statistically significant difference between the study groups with ($p<0.05$). **Conclusion**: Prediction in classifying the cost prediction in health insurance shows that the Decision Tree appears to generate better accuracy than the cost prediction Logistic Regression algorithm.

Keywords: Health care, Decision Tree, Novel Logistic Regression, Machine Learning, Health Insurance

INTRODUCTION

The purpose of the research is to improve the accuracy of cost prediction in health insurance using novel categorization with machine learning algorithms (Rajczi 2019)**.** It preimates the sequence of predicted health cost insurance using decision trees over novel logistic regression. It is found to be important in today's world since health insurance is more convenient for the patients with respect to various scenarios (Quadagno 2006) a widespread statistical phenomenon with potentially serious implications for health care. It can result in wrongly concluding that an effect is due to treatment when it is due to chance. Ignorance of the problem will lead to errors in decision making (Muremyi et al. 2020; Baum et al. 2021; Pauly 2015). Doctors observe the views of different patients to make decisions. The applications of predicting health care costs for each patient with high certainty, problems such as accountability could be solved, enabling control over all parties involved in patients care. It could also be used for other applications such as risk assessment in the health insurance business, allowing competitive premium charges, or for the application of new policies by governments to improve public health (Rajczi 2019; Quadagno 2006).

In distinguishing and forecasting cost prediction in health care using novel ranking by comparing machine learning algorithm 880 journals from IEEE Xplore digital library, 480 articles from ScienceDirect, 1430 articles from google scholar and 498 articles from Decision Tree and novel Logistic Regression algorithm are two highly correlated parts in prediction of cost. In this project is to create a model based on statistically significant factors independent variables which will affect premium charges dependent variables by an insurance company (Qudsi 2015). Among all the articles and journals the most cited paper is (Kharrazi et al. 2021) nearly 1340 people cited these articles and this article is useful for future research. An application is also built which can be interlinked with the decision tree over novel logistic regression model to view the result on UI based on the input parameters (Burns 2015). While the health care law in any country does have some rules for companies to follow to determine premiums concerning the value of insurance in the lives of individuals, it becomes important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it. The model is trained on insurance data from the past (Nakamura et al. 2021; Rushing 1986). The requisite factors to measure the payments can then be defined as the model inputs, then the model can correctly anticipate insurance. We use a decision tree in this analysis since there are many independent variables used to calculate the dependent (target) variable. For this study, the dataset for cost of health insurance is used preprocessing of the datasets (Cebul et al. 2008).(Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020)

Previously our team has a rich experience in working on various research projects across multiple disciplines(Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020).The methods which were used before have less accuracy and detection rate in predicting cost in health insurance (Shiny Irene et al. 2021). Determine premiums,it's really up to know the most important factors and how much statistical importance they hold. In order to sequence the methods and techniques in this research study, Decision Tree generally fares better than novel Logistic Regression (Rajczi 2019). It also takes more time to train a Decision Tree approach for analysing health Insurance datasets which are based on the model to increase with the size of the datasets. Its estimates and calculations are probably done using a characteristic technique. Applying Decision Tree over Logistic Regression model to Medical Insurance dataset to predict future insurance costs for the individuals (Willemse-Duijmelinck et al. 2017). Machine learning is a method of data analysis which sends instructions(programmable code) to code

to computers so that they can learn from data (Cabrera 2011). Then, based on the learned data,they provide us with the predicted results/patterns. The aim of the research work is to predict the cost in health insurance using machine learning and comparing ML techniques Decision Tree over Logistic Regression.

MATERIALS AND METHODS

The research work was performed in the Image Processing Lab, Department of computer Science and Engineering , Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. Basically it is considered that two groups of classifiers are used, namely Decision Tree and Logistic Regression algorithms, which are used to predict cost in health insurance. Group 1 is the Decision Tree with the sample size of 10 and the Logistic Regression is Group 2 with sample size of 10 and they are compared for more Accuracy score and Precision score values for choosing the best algorithm. The Pre-test analysis has been prepared by having a G power of 92% and threshold 0.05%, CI 95% mean and standard deviation (Muremyi et al. 2020). Sample size has been calculated and it is identified that 10 samples/ group in total 20 samples with a standard deviation for Decision Tree = 94.3% and Logistic Regression = 92.7%.Sample size has been calculated and it is identified as standard deviation for Decision Tree= 0.28270 and Logistic Regression = 0.21107 .

A minimum of 4GB ram is required to use and install all the necessary applications, and a minimum of i3 processor to run all the applications and processes simultaneously. A minimum of 50GB hard disk space is required to install the required software and files, and an internet connection is required to download and install all the necessary software and development environment to run this Novel Effective framework. Python programming language is used for the application. The version of python used is 3.9, and the IDLE is used to run and execute the application.

Decision Tree

Decision tree modeling is done by constructing a decision tree for the entire input data. In the Decision tree input 10 variables are represented in the branches and output values are represented in the leaves. Decision tree is one of the predictive modeling approaches used in statistics, data mining and machine learning. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

Pseudocode for Decision Tree

```
Import pandas as pd
Import Matplotlib.pyplot as plt
Import DecisionTreeClassifier
import Decision Tree as Dt
compare from sklearn.tree import DecisionTreeClassifier
skip from sklearn.linear_model import Decision Tree
Data extraction from sklearn import svm
initiate sklearn.metrics import accuracy score
calculate sequence from sklearn.model_selection import train_test_split
Find results from sklearn.feature_extraction.text import CountVectorizer
count_vectorizer=CountVectorizer()
```

```
cv=count_vectorizer.fit()
cv.shape
Dt=Random forest()
Dt.fit(X_train,Y_train)
prediction_Dt=Dt.predict(X_test)
Print(accuracy_score(prediction_Dt
actiom_dt,Y_test))
```

Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as health Insurance cost prediction. In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1

Pseudocode for Logistic Regression

```
Import Logistic Regression classifier
import images
Input file path [filename, pathname] =
uigetfile({'*.jpg'; '*.bmp'; '*.tif'; '*.gif';
'*.png'; '*.jpeg'},
'Load Image File'
if isequal(filename,0)||isequal(pathname,0)
warndlg('Press OK to continue',
'Warning');
else
image aqa = imread([pathname filename]);
imshow(image aqa);
title('Input');
[ image aqa ] = Preprocess( image aqa );
imshow(image aqa);
title('Preprocess');
image aqa = imresize(image aqa);
Resize and reshape the images and form a
cluster pixel;
Get the infection image clearly by
increasing the  pixel from iteration to
iteration;
Plot the graph for accuracy;
Plot  the graph for specificity;
Accuracy of the Logistic Regression
classifier;
```

Statistical Analysis

The analysis was done using IBM SPSS version 21. It is a statistical software tool used for data analysis. For both proposed and existing algorithms 10 iterations were done with a maximum of 15 samples and for each iteration the predicted accuracy was noted for analyzing accuracy. There is a statistically significant difference between the study groups with ($p<0.05$). The value obtained from the iterations of the Independent Sample T-test was performed. The independent data sets are groups,accuracy and details. The Dependent values are patient data and cost. A detailed analysis has been done on these values for finding health insurance in machine learning.

RESULTS

The dataset is provided which selects the random samples from a given dataset for health insurance. The data is collected for a period of 10 days. Group Statistics of decision tree with logistic regression by grouping the iterations with

sample size 10,Mean = 0.92.Group Statistics of Decision Tree with Logistic Regression by grouping the iterations with Sample size 10, Mean = 94.300 , Standard Deviation = 2.02430, Standard Error Mean = 0.64014. The data is collected for a period of 10 days.The model efficiently analyzes health insurance. The mean difference, standard deviation difference and significant difference of Decision Tree based cost prediction and Logistic Regression based cost prediction is tabulated in Table 3,which shows there is a significant difference between the two groups since $p<0.001$ with an independent sample T-Test.

Table 1Health Insurance data set with five attributes which selects the random samples from a given dataset. Table 2 the dataset is the independent and dependent variable.Fig. 1 represents the comparison of Decision Tree over Logistic Regression in terms of mean accuracy. The dependent variables in Health Insurance cost prediction are predicted with the help of the independent variables. The statistical analysis of two independent groups shows that the Decision Tree has a higher accuracy mean (94.3%) compared to the Logistic Regression.

DISCUSSION

In this research clustering has proven to be a highly effective and versatile approach for prediction (Kharrazi et al. 2021; Muremyi et al. 2020). The applied unsupervised learning with R explains clustering methods,distribution analysis,data encoders,and features of R that enable us to to understand the given data better and get answers to our point of prediction problems. The most important features of detecting cost in health insurance using Decision Tree is empirically proven to be a highly effective than Logistic Regression classifier.It was observed that the important factors and the concept of health insurance (Muremyi et al. 2020; Baum et al. 2021)

It was observed that the important factors and the concept of cost prediction has been analysed based on the matrix for health insurance (Willemse-Duijmelinck et al. 2017). The discussion offers insights into the reasons for the development of the practical approach, a concrete methodology for its implementation and strategic and tactical applications of the cost prediction (Maharani et al. 2019). In this the decision tree is also used for predicting the output of the given data. Thus the decision tree shows the more accurate results in it (Daniel 2015) various factors which affect the predicted amount were examined. It was observed that a person's age and smoking status affects the prediction most in every algorithm applied. Attributes which had no effect on the prediction were removed from the features. The effect of various independent variables on the premium amount was also checked. The attributes also in combination were checked for better accuracy results. Premium amount prediction focuses on a person's own health rather than other companies' insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amounts (Rice and Thorpe 1993).

Hence the study results produce clarity in performance with both

experimental and statistical analysis, but it has some limitations to the proposed work such as threshold and precision. The accuracy level of cost prediction in health care using novel based techniques can still be improved by implementing artificial intelligence techniques to predict and analyze results while comparing with existing machine learning algorithms. In the future, the large dataset for health insurance can be considered to validate our proposed model with respect to recent scenarios (Daniel 2015).

CONCLUSION

The current study focused on machine learning algorithms, Decision Tree over Logistic Regression for higher classification in cost prediction. It can be slightly improved based on the Random Forest data sets analysis in future. The outcome of the study shows Decision Tree 94.3% higher accuracy than Logistic Regression 92.7%.

DECLARATIONS

Conflict of Interests

No conflict of interests in this manuscript

Authors Contribution

Author G.Satya Mounika Kalyani was involved in data collection, data analysis, manuscript writing. Author D.Vinod was involved in the Action process, Data verification and validation, and Critical review of manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this research work successfully.

Funding: We thank the following organization for providing financial support that enabled us to complete the study.

1. SNEW.AI Technologies, Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

1. Baum, Griffin R., Geoffrey Stricsek, Mathu A. Kumarasamy, Vineeth Thirunavu, Gregory J. Esper, Scott D. Boden, and Daniel Refai. 2021. “Current Procedural Terminology-Based Procedure Categorization Enhances Cost Prediction of Medicare Severity Diagnosis Related Group in Spine Surgery.” *Spine* 46 (6): 391–400.
2. Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. “A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres.” *Materials Today: Proceedings* 33 (January): 798–805.
3. Burns, Sarah K. 2015. “Health Care Analysis for the MCRMC Insurance Cost Model.” https://doi.org/10.21236/ada618075.
4. Cabrera, Delia. 2011. “A Review of Algorithms for Segmentation of Retinal Image Data Using Optical Coherence Tomography.” *Image Segmentation*. https://doi.org/10.5772/15833.
5. Cebul, Randall, James Rebitzer, Lowell Taylor, and Mark Votruba. 2008. “Unhealthy Insurance Markets: Search Frictions and the Cost and

Quality of Health Insurance." https://doi.org/10.3386/w14455.

6. Daniel, Ines. 2015. "MARKET SEGMENTATION USING COLOR INFORMATION OF IMAGES." *International Journal of Electronic Commerce Studies*. https://doi.org/10.7903/ijecs.1400.
7. Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. "Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition." *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.
8. Kharrazi, Hadi, Xiaomeng Ma, Hsien-Yen Chang, Thomas M. Richards, and Changmi Jung. 2021. "Comparing the Predictive Effects of Patient Medication Adherence Indices in Electronic Health Record and Claims-Based Risk Stratification Models." *Population Health Management*, February. https://doi.org/10.1089/pop.2020.0306.
9. Maharani, Dian, The authors are with Universitas Indonesia, Indonesia, Hendri Murfi, and Yudi Satria. 2019. "Performance of Deep Neural Network for Tabular Data — A Case Study of Loss Cost Prediction in Fire Insurance." *International Journal of Machine Learning and Computing*. https://doi.org/10.18178/ijmlc.2019.9.6.866.
10. Muremyi, Roger, Dominique Haughton, Ignace Kabano, and François Niragire. 2020. "Prediction of out-of-Pocket Health Expenditures in Rwanda Using Machine Learning Techniques." *The Pan African Medical Journal* 37 (December): 357.
11. Nakamura, Haruyo, Floriano Amimo, Siyan Yi, Sovannary Tuot, Tomoya Yoshida, Makoto Tobe, Md Mizanur Rahman, Daisuke Yoneoka, Aya Ishizuka, and Shuhei Nomura. 2021. "Developing and Validating Regression Models for Predicting Household Consumption to Introduce an Equitable and Sustainable Health Insurance System in Cambodia." *The International Journal of Health Planning and Management*, July. https://doi.org/10.1002/hpm.3269.
12. Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. "Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia." *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.
13. Pauly, Mark. 2015. "Cost-Effectiveness Analysis and Insurance Coverage: Solving a Puzzle." *Health Economics*. https://doi.org/10.1002/hec.3044.
14. Quadagno, Jill. 2006. "Cost Containment versus National Health Insurance." *One Nation, UninsuredWhy the U.S. Has No National Health Insurance*. https://doi.org/10.1093/acprof:oso/9780195160390.003.0006.
15. Qudsi, Dini Hidayatul. 2015. *Predictive Data Mining of Chronic Diseases Using Decision Tree: A Case Study of Health Insurance Company in Indonesia*.
16. Rajczi, Alex. 2019. "Personal Cost." *The Ethics of Universal Health Insurance*. https://doi.org/10.1093/oso/9780190946838.003.0004.
17. Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy

Velumurugan. 2020. "Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City." *Oral Health & Preventive Dentistry* 18 (1): 379–86.

18. Rice, Thomas, and Kenneth E. Thorpe. 1993. "Income-Related Cost Sharing in Health Insurance." *Health Affairs*. https://doi.org/10.1377/hlthaff.12.1.21.
19. Rushing, William A. 1986. "Health Insurance, Cost Containment, and Social Conflict: A Future Perspective." *Social Functions and Economic Aspects of Health Insurance*. https://doi.org/10.1007/978-94-009-4231-8_9.
20. Sathish, T., and S. Karthick. 2020. "Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite." *Materials Today: Proceedings* 33 (January): 2555–58.
21. Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. "Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid." In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.
22. Shiny Irene, D., V. Surya, D. Kavitha, R. Shankar, and S. John Justin Thangaraj. 2021. "An Intellectual Methodology for Secure Health Record Mining and Risk Forecasting Using Clustering and Graph-Based Classification." *Journal of Circuits Systems and Computers* 30 (08): 2150135.
23. Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. "Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies." *Materials Today Communications* 25 (December): 101603.
24. Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.
25. Willemse-Duijmelinck, Daniëlle M. I. D., Daniëlle M. I. Willemse-Duijmelinck, Wynand P. M. M. van de Ven, and Ilaria Mosca. 2017. "Supplementary Insurance as a Switching Cost for Basic Health Insurance: Empirical Results from the Netherlands." *Health Policy*. https://doi.org/10.1016/j.healthpol.2017.08.003.

TABLES AND FIGURES

Table 1. Health Insurance datasets with seven attributes which selects the random samples from a given dataset. Implement Decision Tree and Logistic Regression for each data point.

Age	Sex	Bmi	Children	Smoker	Region	Charges
18	female	30.115	0	no	North east	21344.8467

61	female	29.92	3	yes	South east	30942.1918
34	female	27.5	1	no	South west	5003.853
20	male	28.025	1	yes	North west	17560.37975
19	female	28.4	1	no	South west	2331.519

Table 2. Group Statistics of Decision Tree with Logistic Regression by grouping the iterations with Sample size 10, Mean = 94.300 , Standard Deviation = 2.02430 , Standard Error Mean = 0.64014. Descriptive Independent Sample Test of Accuracy and Precision is applied for the dataset in SPSS. Here it specifies Equal variances with and without assuming a T-Test Score of two groups with each sample size of 10.

	Group	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
Accuracy	DT	10	94.3000	2.02430	0.64014
	LOR	10	92.7000	1.92642	0.60919
Loss	DT	10	1.4990	0.28270	0.08940
	LOR	10	2.3920	0.21107	0.6675

Table 3. Independent Sample Test of Accuracy and Precision (calculate P-value <0.001 and Significant value= 0.864, Mean Difference= 1.600 and confidence interval = (0.88-0.11).Decision Tree and Logistic Regression are significantly different from each other.

		Levene's Test for Equality of Variance		**t - test for Equality of Means**		**Sig (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% Confidence Interval of the**	
		F	**Sig**	**t**	**df**				**Lower**	**Upper**
Accuracy	Equal Variance Assumed	0.030	0.864	1.811	18	< .001	1.60000	0.88368	-0.25654	3..45654
	Equal Variance Not Assumed			1.811	17.95	< .001	1.60000	0.88368	-0.2568	3.45687
	Equal	1.030	0.324	-8.004	18	< .001	-	0.111	-	-

Loss	Variance Assumed						0.89300	57	1.12739	0.65861
	Equal Variance Not Assumed			-8.004	16.65	< .001	-0.89300	0.11157	-1.12876	-0.65724

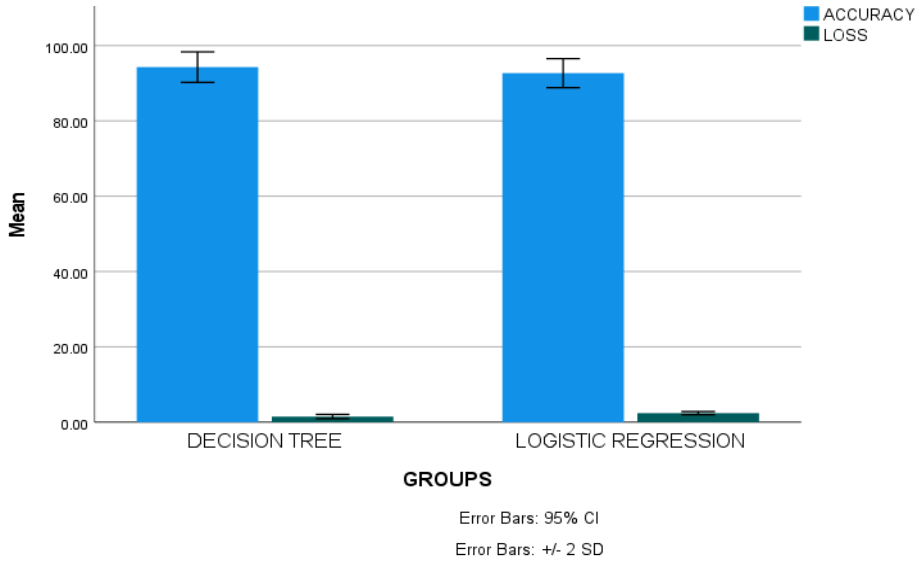

Fig. 1.. Comparison of Decision Tree over Logistic Regression in terms of mean accuracy.It explores that the mean accuracy is slightly better than logistic regression and the standard deviation is moderately improved compared to logistic regression.Graphical representation of the bar graph is plotted using group id as X-axis Decision Tree vs Logistic Regression,Y-Axis displaying the error bars with a mean accuracy of detection +/- 2 SD.