By contrasting linear regression with decision trees, a formal method to predict health insurance costs using a novel symptomatic approach is shown.

G.Satya Mounika Kalyani [1], Senthil Kumar.R [2]*

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, India, Pin Code: 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 research aims to classify the cost prediction in health insurance using novel symptomatic approaches with machine learning algorithms. **Materials and Methods:** The categorizing is performed by adopting a sample size of N=10 in Linear and sample size N=10 in decision tree 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 and Discussion:** The analysis of the results shows that the linear regression has a high accuracy of (95.20) in comparison with the decision tree algorithm (94.30). 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 Linear Regression appears to generate better accuracy than the cost prediction decision tree algorithm.

Keywords: Healthcare, Medical Price Prediction System, Decision Tree, Novel Symptomatic Approach, Machine Learning, Linear Regression.

INTRODUCTION

The purpose of the research is to improve the accuracy of cost prediction in health insurance using novel symptoms with machine learning algorithms (Rajczi 2019). It preimates the sequence of predicted health cost insurance using linear regression over decision trees. 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 Linear regression and Decision Tree 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 linear regression over decision tree 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 multiple linear regression 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).(Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 2020)

The methods which were used before have less accuracy and detection rate in predicting cost in health insurance. 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 Linear Regression generally fares better than Decision Tree (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 Linear regression over Decision Tree 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 research aims to classify the cost prediction in health insurance using novel symptomatic approaches with machine learning algorithms.

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 Linear Regression and Decision Tree algorithms, which are used to predict cost in health insurance. Group 1 is the Linear Regression with the sample

size of 10 and the Decision Tree 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 94% 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 Linear Regression = 95.25% and Decision Tree = 94.3%.Sample size has been calculated and it is identified as standard deviation for Linear regression = 0.28270 and Decision Tree = 0.21306 .

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.

Linear Regression

Multiple Linear Regression is a basic, machine regression model,in which there is one dependent variable and multiple independent variables. The value of a dependent variable is miscalculated from independent variables. The dependent variable is medical charges and independent variables are age,gender,smoker,BMI,children,region.
Multiple Linear Regression uses the ordinary least squares method to find a best fitting line,which involves multiple independent variables.

Pseudocode for Linear Regression

Import Linear Regression as LR
Import pandas as pd
ImportMatplotlib.pyplot as plt
Compare from sklearn ensemble import Linearregression
From sklearn .tree import DecisionTreeClassifier
Data extraction from sklearn
Initiate sklearn.metrics import accuracy price
Calculate sequence sklearn.mode_selection import train_test_split
Find results from sklearn.feature_extraction.value import CountVectorizer
count_vectorizer=CountVectorizer()
Cv.Shape

lrg=LRclassifier(max_depth=8,estimators =100,nthread=3)
lrgc.fit(x_train,y_train)
prediction_lrg=lrgc.predict(x_test)

print(accuracy_price(prediction_lrg,y_test)

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. The 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))
```

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 10 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 linear regression with decision tree by grouping the iterations with sample size 10,Mean = 0.95.Group Statistics of Linear Regression with Decision Tree by grouping the iterations with Sample size 10, Mean = 95.2000 , Standard Deviation = 1.51309, Standard Error Mean = 0.47848. The data is collected for a period of 10 days.The model efficiently to analyze health insurance.The mean difference, standard deviation difference and significant difference of Linear Regression based cost prediction and Decision Tree 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.Figure 1 represents the comparison of Linear Regression over Decision Tree 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 Linear Regression has a higher accuracy mean (95.2%) compared to the Decision Tree.

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 Linear Regression is empirically proven to be a highly effective than Decision Tree 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 analyzed 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 linear regression is also used for predicting the output of the given data. Thus the linear regression 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, Linear Regression over Decision Tree for higher classification in cost prediction. It can be slightly improved based on the Random Forest data sets analysis in the future. The outcome of the study shows Linear Regression (95.20%) higher accuracy than Decision Tree (94.30%).

DECLARATIONS

Conflict of Interests

No conflict of interests in this manuscript

Authors Contribution

Author GSMK was involved in data collection, data analysis, manuscript writing. Author DV was involved in the Action process, Data verification and validation, and Critical review of manuscript.

Acknowledgment

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. Burns, Sarah K. 2015. "Health Care Analysis for the MCRMC Insurance Cost Model." https://doi.org/10.21236/ada618075.
3. 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.
4. 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.
5. Daniel, Ines. 2015. "MARKET SEGMENTATION USING COLOR INFORMATION OF IMAGES." *International Journal of Electronic Commerce Studies*. https://doi.org/10.7903/ijecs.1400.
6. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
7. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.
8. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
9. 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.
10. 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.

11. 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.
12. 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.
13. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
14. Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
15. Pauly, Mark. 2015. "Cost-Effectiveness Analysis and Insurance Coverage: Solving a Puzzle." *Health Economics*. https://doi.org/10.1002/hec.3044.
16. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1). https://doi.org/10.1186/s12302-021-00501-2.
17. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
18. 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.
19. Qudsi, Dini Hidayatul. 2015. *Predictive Data Mining of Chronic Diseases Using Decision Tree: A Case Study of Health Insurance Company in Indonesia*.
20. Rajczi, Alex. 2019. "Personal Cost." *The Ethics of Universal Health Insurance*. https://doi.org/10.1093/oso/9780190946838.003.0004.
21. 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.
22. 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.

23. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
24. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." *Materials Today: Proceedings* 45 (January): 5759–63.
25. Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." *Journal of Circuits Systems and Computers* 30 (05): 2130006.
26. 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 Linear Regression and Decision Tree for each data point.

Age	Sex	Bmi	Children	Smoker	Region	Charges
57	Female	31.16	0	Yes	North west	43578.939
56	Female	27.20	0	No	South west	11073.176
46	Female	27.74	0	No	North west	8026.6666
55	Female	26.98	0	No	North west	11082.577
21	Female	39.49	0	No	South east	2026.9741

Table 2. Group Statistics of Linear Regression with Decision Tree by grouping the iterations with Sample size 10, Mean = 95.2500 , Standard Deviation = 1.51309 , Standard Error Mean = 0.47848. 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	LR	10	95.2500	1.51309	0.47848

	DT	10	94.3000	2.02430	0.64014
Loss	LR	10	1.4990	0.28270	0.08940
	DT	10	2.3780	0.21306	0.06738

Table 3. Independent Sample Test of Accuracy and Precision (Calculate P-value <0.001 and Significant value= 0.306, Mean Difference= 0.950 and confidence interval = (0.79- 0.11). Logistic Regression and Decision Tree are significantly different from each other.

		Levene's Test for Equality of Variance		**t - test for Equality of Means**		**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	1.109	0.306	1.189	18	0.250	0.95000	0.79920	-0.72906	2.62906
	Equal Variance Not Assumed			1.189	16.664	0.251	0.95000	0.79920	-0.73877	2.63875
Loss	Equal Variance Assumed	0.957	0.341	-7.852	18	0.000	-0.87900	0.11194	-1.11419	-0.64381
	Equal Variance Not Assumed			-7.852	16.730	0.000	-0.87900	0.11194	-1.11547	-0.64253

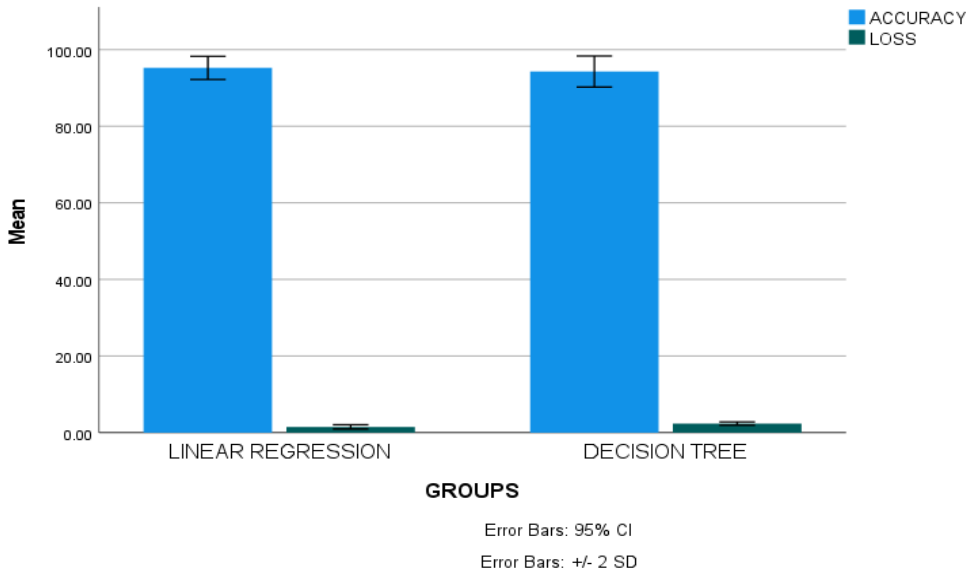

Fig. 1. Comparison of Linear Regression over Random Forest in terms of mean accuracy. It explores that the mean accuracy is slightly better than random forest and the standard deviation is moderately improved compared to random forest. Graphical representation of the bar graph is plotted using group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