Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220070

Comparison of Accuracy Prediction of Medical Insurance Using Decision Tree with K-Nearest Neighbour

AkkarapalliChenchu Krishna[a] and L. Rama Parvathy[b,1]
[a] *Research Scholar, Dept. of CSE, Saveetha School of Engineering,*
[b] *Professor, Dept. of AI&DS, Saveetha School of Engineering,*
[a,b]*SIMATS, Chennai,Tamilnadu, India.*

Abstract: The main aim of this work is to measure and compare the accuracy prediction of medical insurance using a Decision tree with the K-nearest neighbor algorithm. Supervised Machine learning Techniques with innovative Decision Trees (N = 50) and K Nearest Neighbour (KNN) (N = 50) are performed. In this study, 100 photos were utilized, 80% of them being trained and 20% being tested, and the sample size for two groups was computed using G power with a pretest power of 0.8. Compared to Decision Tree and statistical analysis using SPSS software, 100 photos were utilized for group 1 (K-Nearest Neighbour). K-Nearest Neighbour has a mean accuracy of 87.410.224, whereas Decision Tree achieves an accuracy of 82.470.290, with a significant value of 0.297. Based on the execution analysis, the K-Nearest Neighbour approach outperforms the Decision Tree algorithm in terms of accuracy.

Keywords. Innovative Decision Tree, K-Nearest Neighbors, Medical Record, Accuracy Rate, Machine Learning, Medical Cost insurance.

1. Introduction

Health insurance is a form of insurance that covers medical expenditures incurred as a result of an illness. These charges might be connected to hospitalization bills, medication, or medical consultation fees [1]. It is critical to precisely anticipate insurance costs based on people's data, such as age, body mass index, smoking status, etc. [2]. We live in a world fraught with danger and unpredictability. People, homes, businesses, assets, and property are all subject to varying degrees of risk. These threats include the chance of death, illness, and property or financial loss [3, 4]. Many scholars have recently focused on medical cost insurance using machine learning approaches.

The finest portions of people's lives are life and well-being. However, because risks cannot always be avoided, the financial sector has devised a variety of goods to safeguard persons and organizations against them via the use of financial resources to repay them [5][6] [3]. As a result, insurance premiums are never-ending. It is the most acceptable approach for meeting our requirements. Because there are so many

[1]L. Rama Parvathy,Department of AI&DS, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Chennai, India. Email:ramaparvathyl.sse@saveetha.com

independent variables in this research, we utilize multiple rectilinear regression to determine the dependent (target) variable [7]. The dataset for insurance costs is used in this investigation. The dataset was initially preprocessed. Then we used training data to create regression models, which were validated using testing data. Different regression models, including decision trees, k-nearest neighbors, linear regression, and random forest regression, were chosen in this study. The primary objective of this study is to develop a new method of predicting insurance prices.

The fact that these high-performing machine learning algorithms are black-boxed, especially in critical use cases, is a disadvantage. Even though health care cost prediction isn't an essential use case, without an interpretable technique, using patients' personal and clinical information for this topic could result in biased results. As a result, the primary goal of our effort is to improve accuracy-based performance (%) with standard significance value using K-Nearest Neighbours and Decision Tree algorithm in an innovative Prediction of Medical Insurance.

2. Materials and Methods

The research was conducted at SSE, SIMATS, Chennai, in the DBMS Lab of the CSE Department. The dataset was obtained from Kaggle, and the data set was trained using Google CoLab software, with the results being exported to Microsoft Excel. For the creative prediction of medical insurance, input photos are employed for training (80%) and testing (20%) in this framework. The suggested approach is based on python-based computer software that uses picture samples from two groups. The sample size for the two groups is calculated utilizing G power with a pretest power of 0.8 and Kaggle inputs of 100 (50*2). The K-Nearest Neighbors technique is used for group 1 performance analysis, while the Decision Tree approach is used for group 2 performance analysis on the Python platform. The architecture Diagram for Medical Cost Insurance Prediction is given in Figure 1.

The input images were taken from the Kaggle dataset for accuracy analysis. The accuracy is measured, the average values are obtained, and the results are compared to current algorithms. The accuracy is used to determine the performance metric for each sample. For each category, ten samples are collected, and the mean precision is calculated, as shown in Table 1. In statistical analysis, these samples are used to determine the mean, standard deviation, and significance values.

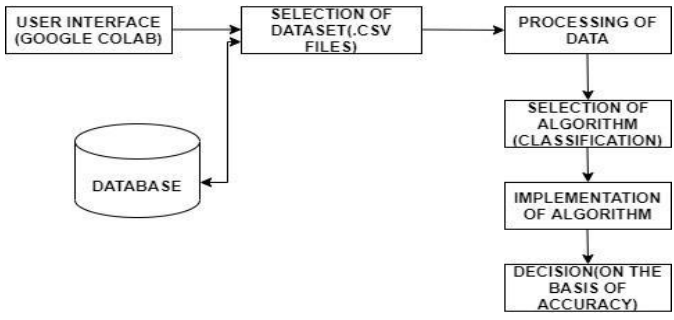

Figure 1. Architecture Diagram for Medical Cost Insurance Prediction

2.1.K-Nearest Neighbours Algorithm

Method of machine learning KNN is a supervised machine learning technique that may be used to solve issues in classification and regression. It is simple to comprehend and put into practice. The K value, a distance metric, is used to classify the data. This parameter will locate new data points and group them according to their attributes and characteristics. In the current system, the K value for categorizing features in medical parameters is set to 7. 80% of the data was utilized for training the system. At the same time, twenty percent was used for testing. Finally, utilizing the current KNN, 72.2% accuracy was obtained for Medical Cost Insurance prediction.

Algorithm for KNN:

Initially assign K value as 7.

KNN(dataset, sample)

1. Calculate the distance between each data item and a certain sample by going through the dataset item by item.
2. Cluster the samples of the organization are generally from the dataset's K-samples with the shortest distance to the sample.
3. Every new item will be classified based on its properties and assigned to existing groups.
4. Repeat step 3 and return the results

2.2 Decision Tree Algorithm

A Decision Tree is a classification and supervised machine technique where the main attribute will be placed as a root node and continues to divide the root node into branches. Branching will continue in both left and right nodes of the tree till it reaches the end of all parameters. A split will occur on every tree level based on the yes or no type.

Root Node = Decision Node Leaf Node = Child Node

Once this process completes, a decision will be taken in a classification manner to give output. The actual data will be split into two parts. One part consists of 80% for training, and another with 20% of the testing data set. In Figure. 2, the decision tree process flow has been mentioned.

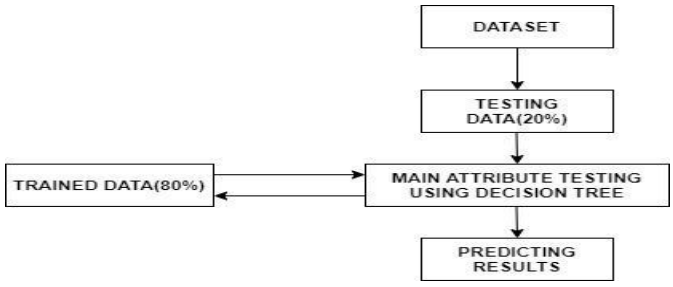

Figure 2. Process diagram Decision Tree Algorithm

Algorithm of Decision Tree

1. Put the best attribute at the top of the tree.
2. Divide the training set into two halves.
3. Repeat both 1 and 2 on both the left and right sides of a tree till you Identify leaf nodes in almost all of the tree's branches.
4. Consider all the branches which reach the leaf node from the root node.
5. Cluster the features using step 4.
6. Repeat from step 2 to step 5 for testing data until clustering into existing groups.
7. Perform accuracy function on major groups and return the value.

Both proposed and existing algorithms are executed in Google Colab, an online platform with python as an integrated development environment and open source to access. The hardware and software specifications include a 64-bit windows system with 4GB RAM and intel core i3 as a processor.

Each dataset is divided into two parts. The first group serves as the training set (80% of the dataset), while the second serves as the testing set (20% of the dataset). Built-in package, which is imported from the sklearn package, will default train the system based on the training dataset. The testing procedure will be based on the classification algorithm we select. This paper will assign 20% of the testing set to the Decision Tree as a parameter. As a result, the testing set will be compared with the Training set and cluster the similar property data points into groups.

We tend to obtain the information set through the Kaggle site to create the claim value model predictor. It includes seven attributes as given in Table 1. The following table depicts the Dataset structure.

Table 1. Random samples from Dataset.

S.no	Age	Sex	BMI	Children	Smoker	Region	Charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.470
5	32	male	28.880	0	no	northwest	3866.855

The data set is divided into two parts: the primary half is known as coaching information, and the second is known as test data; training data accounts for around 80% of the total data utilized, and the remainder is known as test data. The training data set is used to create a model to predict medical insurance costs for the year, and the test data set is used to evaluate the regression model.

Statistical Analysis: An independent sample test in SPSS Software is used for statistical analysis. Descriptive statistical analysis (mean, standard deviation, and expected error mean) were computed for each model. An independent sample T-test, which is used to analyze this study activity, is used to compare the accuracy of the two

groups. The input dataset and epoch time are significant variables, whereas accuracy is a response variable.

3. Results

Each group is executed with ten different datasets at different times to obtain ten sample groups. The IBM SPSS tool will use these sample groups to calculate the independent sample T-Test, which gives significant value for comparing the Decision Tree (DT) and K Nearest Neighbors (KNN) Algorithm. Ten different input datasets are used for comparing both models, and respective accuracy values are recorded below. K-Nearest Neighbors has been classified as a better algorithm than Decision Tree in the prediction of Medical Cost Insurance due to automatic feature interaction in the Dataset.

Table 2 compares two groups using the metrics mean, standard deviation, and standard error mean. It claims that K-Nearest Neighbors outperforms Decision Tree in terms of performance.

Table 2. Statistical Analysis of K Nearest Neighbors& Decision Tree models.

	Group Statistics				
	Group	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	K-Nearest Neighbours	10	87.4100	0.2249	0.7112
	Decision Tree	10	82.4770	0.2905	0.09187

Table 3. Independent sample T-test

		F	**Sig**	**t**	**df**	**Sig (2_tailed)**	**Mean Diff.**	**Std. Error Diff.**	**Lower**	**Upper**
Predicted	Equal Variances assured	1.155	0.297	42.46	18	0.000	4.933	0.1161	4.688	5.177
Actual	Equal Variances not assured			42.46	16.937	0.004	4.933	0.1161	4.687	5.178

Table 3 displays the independent sample. T-tests have a level of significance of 0.000, which is lower than the normal significance range (0.05). As a result, it is demonstrated that Groups 1 and 2 are considerably distinct from one another.

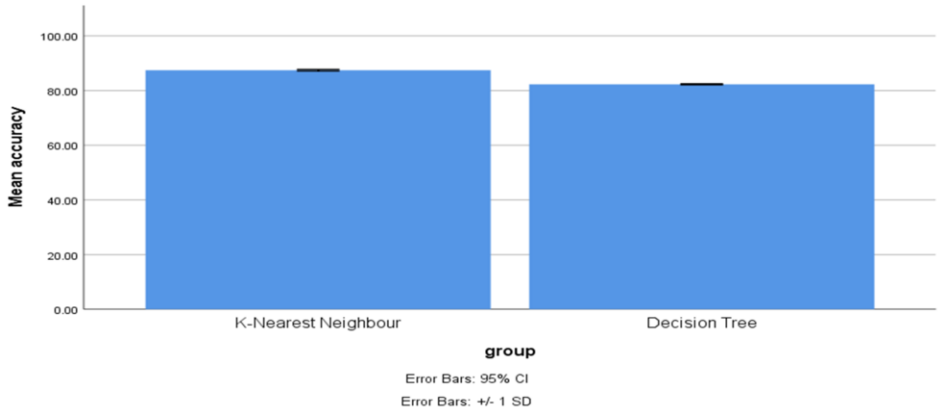

Figure 3. Bar graph between KNN and Decision Tree.

In Figure 3, the Bar chart represents the comparison of Mean Accuracy of Medical Cost Insurance Prediction computed with Decision Tree and KNN algorithms. Decision Tree appears to produce the most consistent results with minimal standard deviation. KNN seems to have the most variable effects with its standard deviation. There is a significant difference between KNN and Decision Tree algorithms ($p< 0.05$ Independent Sample T-Test).

4. Discussion

According to the statistical study, KNN (group 1) has a mean accuracy, variance, and standard error average of 87.4 and 0.7112, respectively. In contrast, the Decision tree (group 2) has a mean accuracy, conflict, and expected error average of 85.06 and 0.9187, respectively. Our overall results obtained by performing SPSS tool calculations state that the proposed KNN (87.4%) algorithm is best suited for Medical Cost Insurance prediction, which satisfies significantly ($p<0.05$) compared to the Decision Tree algorithm (82.7%). But in the proposed model, the results varied with the same instances with decision trees, and the classification led to high accuracy of 98%. [6] used 303 instants with limited attributes using Random Forest and showed an optimistic prediction of 98.45% accuracy with extended classification in the training set, but in the proposed model, less accuracy[8], which is 98% because the collection of decision trees in random forest performs better than single decision trees [8, 9]. The difference, however, was not statically significant ($p<0.05$).

The findings in this paper were almost similar to the above-cited articles. Only the number of parameters and instances of data will decide the accuracy because classification becomes more robust when the data provided in preprocessing is directly related. The accuracy is always dependent on the total number of medical parameters and the number of instants. From the overall literature, many authors have cited proposed better accuracy than existing methods[13]. There are no opposite findings of the above study.

The proposed attempt is limited in that no patients are classified. So one way of improving performance was to classify patients into cost buckets, as recommended by various studies; this led to better performance this season but managed to escape the

goal of this task. For future projects, we can use this categorization process to obtain a clinical risk class as a first phase to enhance efficiency and then compare the proposed algorithm with even more advanced methods to solve the prediction of medical costs.

5. Conclusion

In this research, the Medical Cost Insurance prediction using an innovative Decision Tree is performed with the Medical Cost Insurance dataset found with an accuracy of 87.4% using KNN, which is a more promising result than the existing Decision Tree Algorithm of 82.4%. This proposed model can be used in clinical areas.

References

[1] Xiong L-P, Deng-wen XIA, Ding T. Simulation analysis of medical insurance cost for urban residents in Kunming, China. *Academic Journal of Second Military Medical University* 2011; 31: 741–744.
[2] Ligon JA. The Effect of Health Insurance Cost-Sharing within Episodes of Medical Care. *The Journal of Risk and Insurance* 1993; 60: 105.
[3] Jia J, Song L, Li L. Impact of basic medical insurance fund risk on the health risk assessment of urban residents. *Work*. Epub ahead of print 16 July 2021. DOI: 10.3233/WOR-205352.
[4] Rajczi A. Personal Cost. *The Ethics of Universal Health Insurance* 2019; 81–165.
[5] Takeshima T, Keino S, Aoki R, et al. Development of Medical Cost Prediction Model Based on Statistical Machine Learning Using Health Insurance Claims Data. *Value in Health* 2018; 21: S97.
[6] Li K, Shi Q, Liu S, et al. Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree. *Medicine* 2021; 100: e25813.
[7] Kan HJ, Kharrazi H, Chang H-Y, et al. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PLoS One* 2019; 14: e0213258.
[8] Fordney M. *Fordney's Medical Insurance Dictionary for Billers and Coders - E-Book*. Elsevier Health Sciences, 2011.
[9] Sugita Y. *Japan's Shifting Status in the World and the Development of Japan's Medical Insurance Systems*. Springer, 2018.
[10] Peck G. *Decision Trees, Discriminant Analysis, Logistic Regression, Svm, Ensamble Methods and Knn With Matlab*. 2017.
[11] Luk SCY. *Financing Healthcare in China: Towards universal health insurance*. Routledge, 2016.
[12] Yeh MC-H, Wang Y-H, Yang H-C, et al. Artificial Intelligence-Based Prediction of Lung Cancer Risk Using Nonimaging Electronic Medical Records: Deep Learning Approach. *J Med Internet Res* 2021; 23: e26256.
[13] Cobre A de F, Stremel DP, Noleto GR, et al. Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? *ComputBiol Med* 2021; 134: 104531.