****

HEALTH INSURANCE COST PREDICTION USING MACHINE LEARNING

**1. INTRODUCTION**

**1.1 Overview**

A stethoscope on a piece of paper

Description automatically generated with low confidenceHealth insurance is a necessity nowadays, and almost every individual is linked with a government or private health insurance company. Factors determining the amount of insurance vary from company to company. Also, people in rural areas are unaware of the fact that the government of India provide free health insurance to those below poverty line. It is very complex method and some rural people either buy some private health insurance or do not invest money in health insurance at all. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance.

**1.2 Purpose**

The goal of this project is to allows a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

**2. LITERATURE SURVEY**

**2.1 Existing problem**

Prediction is premature and does not comply with any particular company so it must not be only criteria in selection of a health insurance. Early health insurance amount prediction can help in better contemplation of the amount needed. Where a person can ensure that the amount, he/she is going to opt is justified. Also, it can provide an idea about gaining extra benefits from the health insurance.

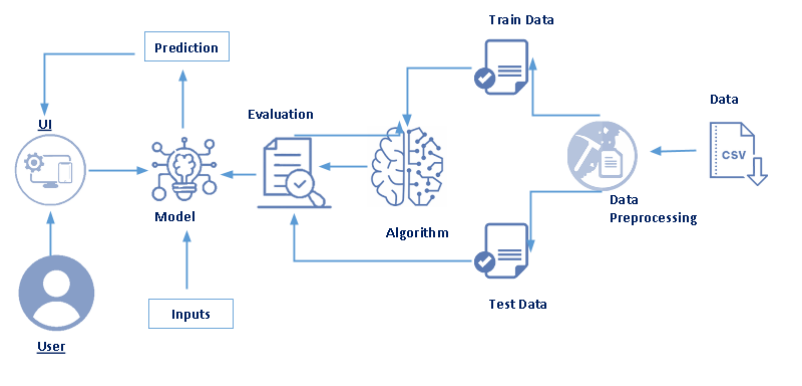
**2.2 Proposed solution**

Early health insurance amount prediction can help in better contemplation of the amount needed. Where a person can ensure that the amount, he/she is going to opt is justified. Also, it can provide an idea about gaining extra benefits from the health insurance.

**3. THERORITICAL ANALYSIS**

**3.1 Block Diagram**

**Architecture:**

****

**3.2 Hardware / Software designing**

**3.2.1 Hardware requirements:**

Processer : Any Update Processer

Ram : Min 4GB

Hard Disk : Min 100GB

**3.2.2 Software requirements:**

Operating System : Windows family

Technology : Python3.7

IDE : Jupiter notebook

Flask : Visual studio

**4.EXPERIMENTAL INVESTIGATIONS**

The primary source of data for this project was from Kaggle user D’marco. The dataset is comprised of 1338 records with 6 attributes. Attributes are as follow ‘age’, ‘gender’, ‘bmi’, ‘children’, ‘smoker’ and ‘charges ‘. The data was in structured format and was stores in a csv file. Dataset is not suited for the regression to take place directly. So, cleaning of dataset becomes important for using the data under various regression algorithms. In a dataset not every attribute has an impact on the prediction. Whereas some attributes even decline the accuracy, so it becomes necessary to remove these attributes from the features of the code. Removing such attributes not only help in improving accuracy but also the overall performance and speed.

Text, table

Description automatically generated

In health insurance many factors such as pre-existing body condition, family medical history, Body Mass Index (BMI), marital status, location, past insurances etc affects the amount. According to our dataset, age and smoking status has the maximum impact on the amount prediction with smoker being the one attribute with maximum effect. Children attribute had almost no effect on the prediction, therefore this attribute was removed from the input to the regression model to support better computation in less time.

## 

## 5. DATA PREPROCESSING

The seven attributes in the figure has some contribution to estimate the cost of the insurance, which is our dependent variable. In this stage, the data is scrutinized and updated properly to efficiently apply the data to the ML algorithms.

First of all, all unknown values are cleaned. The unknown numerical values are replaced with the mean. The target variable (charges) would then be examined .

Because the mean value is greater than the median, this implies that the distribution of health insurance charges is right-skewed. We can confirm this visually using a histogram or density plot .

Table

Description automatically generated

Fig: Descriptive statistic of charges variable

Chart, histogram

Description automatically generated

Fig: Histogram and Density-Plot Diagram

## 

## 5.1 MACHINE LEARNING

Machine learning can be defined as the process of teaching a computer system which allows it to make accurate predictions after the data is fed. However, training has to be done first with the data associated.

It is a type of artificial intelligence ([AI](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence)) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning [algorithms](https://whatis.techtarget.com/definition/algorithm) use historical data as input to predict new output values.

It gives enterprises a view of trends in customer behaviour and business operational patterns, as well as supports the development of new products.

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, [unsupervised](https://whatis.techtarget.com/definition/unsupervised-learning) learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists chooses to use depends on what type of data they want to predict.

Graphical user interface, text, application, chat or text message

Description automatically generated

Fig: Types of Machine Learning

### **5.2 SUPERVISED LEARNING**

Supervised learning algorithms create a mathematical model according to a set of data that contains both the inputs and the desired outputs. Usually, a random part of data is selected from the complete dataset known as training data, or in other words a set of training examples. Training data has one or more inputs and a desired output, called as a supervisory signal. What’s happening in the mathematical model is each training dataset is represented by an array or vector, known as a feature vector. A matrix is used for the representation of training data. Supervised learning algorithms learn from a model containing function that can be used to predict the output from the new inputs through iterative optimization of an objective function. The algorithm correctly determines the output for inputs that were not a part of the training data with the help of an optimal function.

## 

## 5.2 . UNSUPERVISED LEARNING

In this learning, algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. Test data that has not been labelled, classified, or categorized helps the algorithm to learn from it. What actually happens is unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. The main application of unsupervised learning is density estimation in statistics. Though unsupervised learning encompasses other domains involving summarizing and explaining data features also.

## 5.3 REINFORCEMENT LEARNING

Reinforcement learning is class of machine learning which is concerned with how software agents ought to make actions in an environment. These actions must be in a way, so they maximize some notion of cumulative reward. Reinforcement learning is getting very common in nowadays, therefore this field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulated-based optimization, multi-agent systems, swarm intelligence, statistics, and genetic algorithms.

## 

## 6. REGRESSION

The regression analysis is a predictive method that explores the relationship between a dependent (target) and the independent variable(s) (predictor). This technology is used to forecasting, estimate model time series, and find the causal effect relationship among the variables. In this analysis, for example, I want to analyze the relationship between insurance cost (target variable) and six independent variables based on (age, BMI, child number, individual living area, or sex and whether the customer is a smoking person).on the basis of a regression. The regression analysis estimates the relationship between two or more variables, as stated previously. I used different regression models to estimate health insurance costs on the basis of six independent variables, and by using this regression, we can forecast future health insurance fees based on current and past data.

There are several advantages of using regression analysis as follows:

-It demonstrates the essential relationships between the dependent and independent variables.

-It shows the effect intensity on the dependent variable of several independent variables.

Analysis of regression also helps one to compare the results of measured variables at various scales, such as independent variable and dependent variable effects. These advantages allow market researchers, data analysts, and data scientists to remove and determine the best range of variables for predictive model .

# 6.1 REGRESSION MODELS

### **6.1.1.MULTIPLE LINEAR REGRESSION**

Multiple linear regression can be defined as extended simple linear regression. It comes under usage when we want to predict a single output depending upon multiple input or we can say that the predicted value of a variable is based upon the value of two or more different variables. The predicted variable or the variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable) and the variables being used in predict of the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

In practice, we often have more than one predictor. For example, with the data set used in this study, we may wish to understand if independent variables (6 independent variables), (linearly) related to the dependent variable (charges). this is referred to as the multiple linear regression (MLR) model [10]. An MLR model with t\ independent features X1,X2,……Xt and Y results can be calculated as in the following equation

Y= a0X0+a1X1+…………+atXt+u

. In the above equation, u is the residual regression while a is the weight of each independent variable or parameter assigned.

### **6.1.2.DECISION TREE REGRESSION**

Regression or classification models in decision tree regression builds in the form of a tree structure. The dataset is divided or segmented into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision tree with decision nodes and leaf nodes is obtained as a final result. These decision nodes have two or more branches, each representing values for the attribute tested. Decision on the numerical target is represented by leaf node. The topmost decision node corresponds to the best predictor in the tree called root node. Numerical data along with categorical data can be handled by decision tress. DTs are straightforward, very popular , fast-training, and easy to read models with comparative or other methods of learning from the data. They are fairly competent but vulnerable to overfitting in their predictions. They can be strengthened by improving their performance . Different models of DTs, CART, C4.5, etc. are available . and CART (Regression Trees) will be used for this analysis.

### **6.1.3.GRADIENT BOOSTING REGRESSION**

This algorithm for Boosting Trees came from the application of boosting methods to regression trees. The basic idea behind this is to compute a sequence of simple trees, where each successive tree is built for the prediction residuals of the preceding tree. For predictive models, gradient boosting is considered as one of the most powerful techniques. Gradient boosting involves three elements:

1. An optimized loss function.

2. An additive model to add weak learners to minimize the loss function.

3. A weak learner to make predictions

### **6.1.4.GENERALIZED ADDICTIVE MODEL(GAM)**

Generalized additive models are incorporated into the actuary toolkit to deal flexibly with continuous functionality. The continuous features in this setting insert the model into a semi-parametric additive predictor. The impact of the policyholder's age, vehicle power or amount insured may be modelled by GAMs in property and casualty insurance. GAMs also allow actuaries to evaluate geographical risk variances, considering the potential interaction of continuous characteristics. Other experiences in the data usually include age, power and gender, and age in engine insurance. You can also be caught by GAMs.

### **6.1.5.RANDOM FOREST**

Random forests reflect a shift to the bagged decision trees that create a broad number of decorrelated trees so that predictive efficiency can be improved further. They are a very popular 'off-the-box' or off-the-shelf' learning algorithm, with good predictive performance and relatively few hyperparameters. There are several implementations of random forests that exist, but the Leo Breiman algorithm (Breiman 2001) is now largely authoritative. .Random forests create an average predictive value as a result throughout the regression of individual trees. Random forests resolve to overfit . As in the following equation , a random model for forest regressors can be expressed.

g(x)=fo(x)+f1(x)+f2(x)+…….+fn(x)

where g is the final model that is the sum of all models. Each model f(x) is a decision tree.

### **6.1.6.XGBOOST**

A new ensemble learning software named XGBoost has been proposed[13]. Which is a new tree-enhancing model that provides effective out-of-core learning and sparse memory. XGBoost is therefore a supervised learning algorithm, which would be extremely useful for argument prediction issues with broad training data and missing values. Missing values can still not be managed by the most popular approaches, such as random forests and neural networks. Methods require additional frameworks to manage the missing values. The power of XGBoost improves the use of the tool in many other applications. For example, in direct-diffuse solar separation, Aler et al. [14] Developed two versions of XGBoost. The first one is an indirect model, which uses XGBoost to learn solar radiation separation models from various literature sources in a data set from traditional level 1 instruction models. Another model is a direct model that straightforwardly suits XGboost in a dataset. An Additional case is [15], which uses XGBoost to recommend things to a user using functions derived from the pair of users using complicated feature engineering in the recommendation framework. In this study, we analyze XGBoost as a predictor model for the medical insurance cost.

### **6.1.7.SUPPORT VECTOR MACHINE**

SVMs can be generalized to problems with regression (i.e., when the outcome is continuous such as our target variable in our study ). Essentially, SVMs are seeking a hyperplane in an extended function space that usually results in a nonlinear decision limit with strong generalization efficiency in the original feature space. Specific functions called kernel functions are used to build this extended, separated functionality.

### **6.1.8.K NEAREST NEIGHBORS**

K-NN is a very simple predictive model that predicts values on the basis of their "likelihoods" from other values. Contrary to most other machine learning approaches, KNN is dependent on memory and cannot be summed up as a closed algorithm. This implies that the training data are required during operations and forecasts are produced immediately from the training data relations. KNNs are additionally identified as lazy learning and can also be inefficient computationally. Nevertheless, KNNs have succeeded in several market problems.

### **6.1.9.STOCHASTIC GBM’S**

Gradient boosting machines (GBMs) are an extremely common ML regression model. Whereas random forests construct a group of independent deep trees, GBMs construct a set of shallow trees Each tree learns and develops compared to the prior one .

While shallow trees are feeble forecasting predictive models, they can be "boosted" to create a strong committee, which often is difficult to tackle with other algorithms if properly tuned. A significant observation from Breiman was that the training of algorithms on a random training subsample provided more reductions in the tree correlation and thus enhanced predictive accuracy. The same logic was used by Friedman and the boosting algorithm upgraded accordingly. This procedure is called a stochastic gradient boosting.

# 

# 7.DESIGNING AND IMPLEMENTATION

## I.DATA PREPARATION AND CLEANING

The data has been imported from kaggle website. The website provides with a variety of data and the data used for the project is an insurance amount data. The data included various attributes such as age, gender, body mass index, smoker and the charges attribute which will work as the label for the project. The data was in structured format and was stores in a csv file format. The data was imported using pandas library.

The presence of missing, incomplete, or corrupted data leads to wrong results while performing any functions such as count, average, mean etc. These inconsistencies must be removed before doing any analysis on data. The data included some ambiguous values which were needed to be removed.

## II.TRAINING

Once training data is in a suitable form to feed to the model, the training and testing phase of the model can proceed. During the training phase, the primary concern is the model selection. This involves choosing the best modelling approach for the task, or the best parameter settings for a given model. In fact, the term model selection often refers to both of these processes, as, in many cases, various models were tried first and best performing model (with the best performing parameter settings for each model) was selected.

## III.PREDICTION

The model was used to predict the insurance amount which would be spent on their health. The model used the relation between the features and the label to predict the amount. Accuracy defines the degree of correctness of the predicted value of the insurance amount. The model predicted the accuracy of model by using different algorithms, different features and different train test split size. The size of the data used for training of data has a huge impact on the accuracy of data. The larger the train size, the better is the accuracy. The model predicts the premium amount using multiple algorithms and shows the effect of each attribute on the predicted value.

## 

## 8. IMPLEMENTATION

The objective of the study is to predictive the insurance cost based on age, BMI, child number, the region of the person living, sex, and whether a client is smoking or not. These features contribute to our target variable prediction of insurance costs. For the measurement of the cost of insurance, several regression models are applied in this study. The dataset is split into two sections. One part for model training and the other part for model evaluation or testing. In this study, the data set is separated into two-part the first part is called training data and the second called test data, training data makes up about 80 percent of the total data used, and the rest for test data. Every one of these models is trained with the training data part and then evaluated with the test data. For this study, R x64 4.0.2 is used for applying these models. We used two main libraries are CART and Keras for ML and deep learning models. And we used Mean absolute error (MAE), root mean squared error (RMSE) and R-squared As a standard for evaluating these models.

The Mean Absolute Error (MAE) is the difference between the original and forecast values obtained by averaging the absolute difference over the data set.

MAE=1/N

The RMSE of the disparity between the expected values and the real values is determined as the square root. For an accurate forecast, the RMSE must be low so there would be less variance among the expected values and the real values.

RMSE=

Where N=number of overall observations

Yˇ=expected insurance fee values

Y=real insurance fee values

The R-squared is often called the coefficient of decision. The proportion of variance is estimated from the independent variables in the dependent variable.

R-squared= Explained Variance /Total Variance

The more R-squared , the better the model output. , and indicates that the model deviates less from real values. A R-squared score of 1 indicates that it suits perfectly.

To evaluate the performance of various machine learning algorithms (Multiple Linear Regression, Generalized Additive Model, SVM, CART, RF, XGBoost, k-Nearest Neighbours, Stochastic Gradient Boosting, and Deep Neural Network All of these models are trained on the basis of training data and tested on test data. Mean absolute error, RMSE, and R-squared for each of these models are measured. And the given Table displays the results.

Table

Description automatically generatedFig: Model Performance



Chart, bar chart

Description automatically generated

Fig: Comparison Between Models

## 9. RESULT

We see that the accuracy of predicted amount was seen best i.e. 99.5% in gradient boosting decision tree regression. Other two regression models also gave good accuracies about 80% In their prediction. The figure shows the accuracy percentage of various attributes separately and combined over all three models. Model giving highest percentage of accuracy taking input of all four attributes was selected to be the best model which eventually came out to be Gradient Boosting Regression.

Table

Description automatically generated

Fig: Accuracy in percentage(%)

Chart, scatter chart

Description automatically generated

The above figure shows the graphs of every single attribute taken as input to the gradient boosting regression model.

Graphical user interface, text, application

Description automatically generated

A picture containing table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generatedText

Description automatically generated

Graphical user interface, application

Description automatically generatedChart, bar chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedGraphical user interface

Description automatically generated with medium confidenceChart, bar chart

Description automatically generatedGraphical user interface, text, application, chat or text message, email

Description automatically generatedChart, bar chart, treemap chart

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedChart, bar chart

Description automatically generatedGraphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

A picture containing text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated

Graphical user interface, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application

Description automatically generated

# 10. CONCLUSION

The research uses various machine learning regression models and deep neural networks to forecast charges of health insurance based on specific attributes, on medical cost personal data set from Kaggle.com. The findings are summarized in the table shows that Stochastic Gradient Boosting offers the best efficiency, with an RMSE value of 0.380189, an MAE value of 0.17448, and an accuracy of 85.82. Stochastic gradient boosting can therefore be used in the estimation of insurance costs with better performance than other regression models. Forecasting insurance costs based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual. Machine learning can significantly minimize these individual efforts in policymaking, as ML models can do cost calculation in a short time, while a human being would be taking a long time to perform the same task. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data.

**11.FUTURE SCOPE**

In future, this model can be used to compare various machine learning algorithm generated prediction models and the model which will give higher accuracy will be chosen as the prediction model. This paper work can be extended to higher level in future. Predictive model for Heath Insurance that uses machine learning algorithms, where the results from each graph of the paper can be taken as individual criteria for the machine learning algorithm.

**12.BIBILOGRAPHY**

**REFERENCES**

1. L. S. Chen and J. C. Chen, “Using data mining methods to detect medical fraud,” in *Proceedings of the 2020 International Conference on Management of e-Commerce and e-Government*, pp. 89–93, Jeju Island, South Korea, July 2020.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Using%20data%20mining%20methods%20to%20detect%20medical%20fraud&author=L.%20S.%20Chen%20&author=J.%20C.%20Chen)

2. M. A. Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, “Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation,” in *Proceedings of the AMIA Annual Symposium Proceedings*, vol. 2017, American Medical Informatics Association, Washingdon, DC, USA, November 2017.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Supervised%20learning%20methods%20for%20predicting%20healthcare%20costs:%20systematic%20literature%20review%20and%20empirical%20evaluation&author=M.%20A.%20Morid&author=K.%20Kawamoto&author=T.%20Ault&author=J.%20Dorius&author=&author=S.%20Abdelrahman)

3. [https:// en.wikipedia.org/wiki/Exploratory\_data\_analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis)

4. <https://pandas.pydata.org/pandas-docs/stable/>

5. https://smartinternz.com/Student/guided\_project\_info/8835

**APPENDIX**

* **Pandas-** It is a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language.
* **Numpy** - It is an open-source numerical Python library.
* **Matplotlib**- Visualisation with python
* **Sklearn.Preprocessing**- This package provides several common utility functions and transformer classes
* **Seaborn as sns**- Seaborn is a data visualization library in Python based on matplotlib